A CONVOLUTIONAL NEURAL NETWORK APPROACH FOR ABNORMALITY DETECTION IN WIRELESS CAPSULE ENDOSCOPY

Anjany Kumar Sekuboyina[1,], Surya Teja Devarakonda[2,*], and Chandra Sekhar Seelamantula[3]*

[1] Klinikum rechts der Isar der Technische Universität München, München - 81675

[2] Department of Electrical Engineering, Indian Institute of Technology, Hyderabad - 502285

[3] Department of Electrical Engineering, Indian Institute of Science, Bangalore - 560012

email: anjany.sekuboyina@tum.de, ee13b1034@iith.ac.in, chandra.sekhar@ieee.org

ABSTRACT

In wireless capsule endoscopy (WCE), a swallowable miniature optical endoscope is used to transmit color images of the gastrointestinal tract. However, the number of images transmitted is large, taking a significant amount of the medical expert's time to review the scan. In this paper, we propose a technique to automate the abnormality detection in WCE images. We split the image into several patches and extract features pertaining to each block using a convolutional neural network (CNN) to increase their generality while overcoming the drawbacks of manually crafted features. We intend to exploit the importance of color information for the task. Experiments are performed to determine the optimal color space components for feature extraction and classifier design. We obtained an area under receiver-operating-characteristic (ROC) curve of approximately 0.8 on a dataset containing multiple abnormalities.

Index Terms— Gastrointestinal tract, Classification, Wireless capsule endoscopy, Convolutional neural networks

1. INTRODUCTION

A painless method of diagnosing the gastrointestinal tract has been made possible by the concept of Wireless Capsule Endoscopy (WCE), developed by Iddan et al. [1], in which a capsule containing a miniature camera is ingested for the purpose of imaging and simultaneously transmitting the images of the digestive tract through radio-telemetry. This variant of endoscopy offers several advantages, over cable endoscopy, such as reduced patient discomfort, sufficient amount of data owing to the high frame-rate of the capsule endoscope (CE), and a simple scan procedure that does not require the doctor to be present online throughout the scan. However, the number of images transmitted during the journey of the CE are of the order of a few tens of thousands, and it requires considerable amount of focussed offline review by an expert to identify the abnormalities in a single scan. This inconvenience faced by the WCE reviewers can be handled by computer-aided intervention into the review process, thereby making WCE an efficient diagnosis technique.

1.1. Related Literature

Techniques for analyzing an endoscopic video were proposed as early as 2001. Karkanis et al. [2] proposed a neural-network-based approach with textural descriptors extracted from a wavelet transformation. Since then, automated WCE image analysis has been an active area of research with techniques that are either aimed at automatically detecting the presence of some abnormality or aimed at identifying a particular kind of abnormality such as bleeding [3, 4], ulcers [5], polypoids [6], etc. Majority of the approaches analyzing WCE data employed a classifier (support vector machine or neural network) that classifies a descriptive feature from a pixel, thereby identifying if the pixel is part of an abnormality or not. Another domain of WCE video analysis includes reduction of the number of WCE frames that need to be analyzed based on the presence of an abnormality. For example, Iakovidis et al. [7] treated the WCE video as a vector space and obtained a set of orthogonal vectors using non-negative matrix factorization. The orthogonal vectors correspond to the representative video frames of the WCE video, in terms of which the entire video can be summarized.

The features that were used in the classifiers for detection of abnormalities tried to capture color and textural information. Lv et al. [3] employed transformed color histograms as a color invariant descriptor for identifying bleeding. Iakovidis et al. have explored the domain of abnormality detection in great detail and made seminal contributions. In [8, 9], they detected salient points in an image based on a color channel and constructed discriminative features capturing the color information around the salient points based on a feature detection algorithm. In a similar approach, features were extracted from salient superpixels for the detection of blood [4]. From a pure texture analysis perspective, Li et al. [5] proposed a descriptor based on curvelet for a local binary pattern. In another approach [6], they also fused the information from the local binary pattern with the wavelet information to propose a new textural feature for polyp detection. In addition to these two classes of algorithms, namely, generic and specific abnormality detection, there exists a third class that aims at detecting more than one type of abnormality, for example, Karargyris et al. [10] identified ulcers and polyps by combining log Gabor filter-based segmentation, color transformation, and pattern recognition using edge detection. Similarly, Szczypiski et al. [11] combined texture and color information across color spaces in order to identify ulcers and bleeding.

1.2. Our Contribution

Our work is motivated by the research of Iakovidis et al. [8], who take a two-step approach. First, speeded-up robust features (SURF) interest points are extracted from the WCE images. The interest points are then used to extract features for classification. However,

[*] Both authors contributed equally and must be treated as joint first authors. This work was sponsored by the Robert Bosch Centre for Cyberphysical Systems, Indian Institute of Science, Bangalore - 560 012, India.

SURF being a blob-detection algorithm, is not an optimal feature extractor for several abnormalities such as bleeding, stenosis, villous edema, large polypoids, etc., which do lack a distinguishing pattern. In our approach, we do away with SURF, and instead extract features from every pixel in the WCE frames. Another shortcoming of SURF interest points is that the feature vector is crafted manually, by using the color information of the pixel and its variation from the surrounding pixels. This approach does not entirely capture the variation in the textural color patterns of the images, which is an important indicator for many diseases. Additionally, the complexity of the WCE data and the wide range of diseases that are considered make it extremely difficult for hand-crafted features to generalize well across diseases. We address this issue by making use of a CNN to learn the textural color patterns of abnormalities and extract artificial features from various regions of the endoscopic frame. A CNN learns local correlation among pixels and permeates this correlation at various hierarchies resulting in an abstract representation for every pixel neighborhood. This representation is used as a feature in the classification process. Due to its inherent feature generating capabilities, a CNN generalizes well across diseases, provided that it has sufficient training data to learn the representations from. CNNs are used widely in the field of image processing due to their ability to detect spatial patterns in a robust manner [12].

We show that the proposed classifier gives a comparable performance in terms of the area under ROC curve (AUC) with respect to [9]. In Section 2, we provide an overview of the classification algorithm. Section 3 describes the architecture of the CNN, various data processing techniques that have been studied, and the performance of the base CNN. We consolidate our conclusions in Section 4.

1.3. Description of the WCE Dataset

The WCE dataset was originally recorded and shared by Iakovidis et al. [8]. For the sake of completion, we provide a description of the dataset. The CE captures images at a rate of three frames per second with a spatial resolution of 320×320 pixels. The video frames captured by the CE were classified by experts into broad categories such as inflammatory lesions, vascular lesions, lymphangiectasias, and polypoid lesions. A subset of 137 images from the WCE dataset was used for experimentation, of which 77 images had abnormalities. These images were annotated at a pixel-level by experts, and the annotated pixels were used as the class labels during classification. The composition of the dataset is as follows: 5 images each of aphthae and intraluminal bleeding, 9 images of nodular lymphangiectasias, 8 images of chylous cysts, 27 images of angiectasias, 6 images each of polyps and stenoses, 9 images of ulcers, and 2 images of villous oedema, and 60 normal frames.

2. METHODOLOGY

Texture and color are the two characteristic features that play a distinct role in detecting abnormalities in an image. The significance of texture for abnormality detection can be observed from its use in [5] and [11]. Further, an investigation of the WCE database gives sufficient ground for using features based on color information. From bleeding to polypoids, ulcers and stenosis, a clear contrast in color is observable with respect to the surrounding region. The importance of color in the detection of lesions was also examined in [8], [9], and [4]. From these studies, we draw our motivation for generating features based on texture and color information.

In this work, we exploit the color information by using the chromatic components of the images, with the components taken from various color spaces that have been previously considered in WCE video analysis, which include CIE-*Lab* (using standard illuminant D65) and YCbCr spaces [8], and the texture information by using a CNN to learn the textural patterns from the chromatic components.

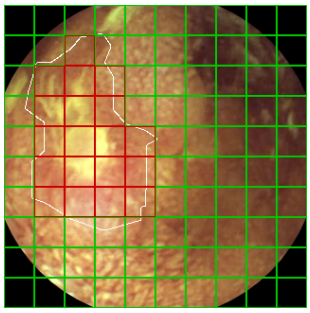


Fig. 1: Patch extraction: The patches in red have more than 50% malignant pixels and are labeled as *malign*. The remaining patches in green are labeled as *benign*. The contour in white denotes the boundary of the malignant pixels, as given in the manually annotated ground truth.

Our algorithm aims at classifying a pixel in the frame as one belonging to an abnormality (*malign* pixel) or not (*benign* pixel). It proceeds as described below:

1. **Data processing**: Transform the color space of the endoscopic image to one of the following color spaces: CIE-*Lab* and YCbCr.
2. **Training**: Extract a feature vector from every pixel using a CNN. We feed pixel-neighborhoods (patches) to the CNN instead of the entire image. This is because the lesions can occur anywhere in the image, and we intend to learn only the appearance and not their location. Further, considering the neighborhoods of all the pixels results in a lot of redundant information, thereby causing an over-fit. Thus, we split the endoscopic frame into non-overlapping patches, each labelled as a *benign* or *malign* patch based on the percentage of malignant pixels in the patch (in our case, if more than 50% of the pixels in a patch are labelled *malign*, the patch is considered to be *malign*), as shown in Fig. 1.
3. **Inference**: Classify the patches using the trained model. After the classification, each pixel is assigned the label of the patch to which it belongs.

3. TRAINING AND INFERENCE

In this section, we describe the data pre-processing that was needed, the experiments that helped identify the optimal color component to be used, the training procedure that was employed, and finally the optimal CNN architecture. About 90% of the randomly sampled frames from the WCE dataset constituted the training set and the remaining frames were used interchangeably for validation and testing. The performance of the network was evaluated using measures such as sensitivity (SN), specificity (SP), and the AUC.

3.1. Data Imbalance Mitigation

The WCE dataset is highly imbalanced towards *benign* data. If the number of *benign* examples are much larger than the number of *malign* examples, the classifier will tend to classify the *malign* examples

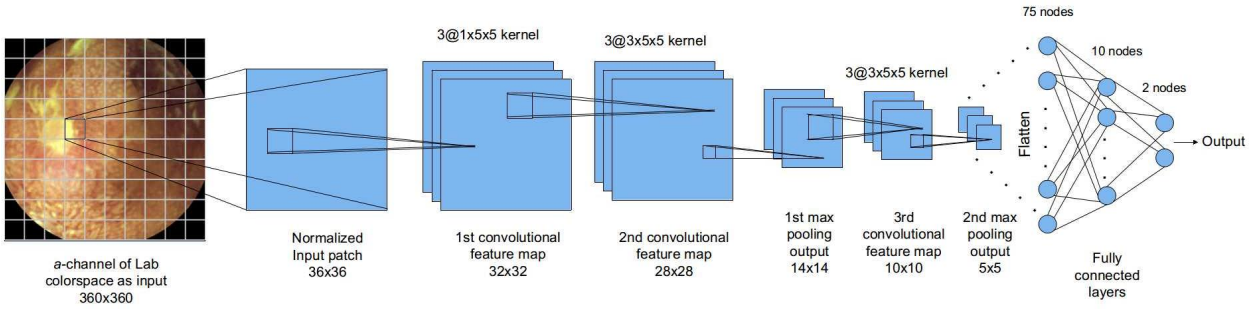


Fig. 2: Optimal CNN architecture used for the classification task. Detailed description of the network is given in Section 3.3.

Table 1: A comparison of the SN and SP (SN: sensitivity; SP: specificity) values obtained on both training and testing datasets by using various data expansion methods during the training phase.

Method	Training Data		Testing Data	
	SN (in %)	SP (in %)	SN (in %)	SP (in %)
Oversampling	97	95	47	92
Reflection	98	84	70	91
SMOTE	96	92	79	87

also as *benign* ones, as that results in an overall high accuracy. This problem can be overcome by using sampling-based approaches that oversample the minority class data points, at the patch level, to reduce the impact of data imbalance. The *a*-channel of the CIE-*Lab* color space (the reason for this choice is elaborated in Section 3.2) is used to test various oversampling techniques. The oversampling techniques considered are described below :

1. *Oversampling* : In this approach, each of the minority class data points is reused multiple times. This method is fast and simple. However, the repetition of data affects the generality of the network and leads to over-fitting.
2. *Reflection* : In this approach, random translation and rotation of data points is used to generate extra samples. In WCE data acquisition, the images are affected by the capsule's continuous rotatory and translatory motion. Data expansion using *reflection* helps in accounting for these situations, thus improving the generality of the model.
3. Synthetic minority oversampling: In this oversampling technique, also referred to as SMOTE [13], extra samples are generated as interpolations between minority class data points and their k minority class nearest neighbors, which are found using the structural similarity (SSIM) index. This method also improves generality and prevents over-fitting.

To validate the optimal data expansion method for our experiment, a CNN with one convolutional layer with three kernels having a planar spread of 5×5 and one max pooling layer with a pool size 2×2, followed by node flattening and a multi-layer perceptron network of structure 50-10-2 is used for the classification task. The results obtained are presented in Table 1. We observe that *SMOTE* and *reflection* techniques perform better than *oversampling*. Further, synthetic data generation using *reflection* is faster than *SMOTE*, which involves computationally intensive nearest-neighbour search. Hence, the *reflection* technique is used for data expansion of minority class data during the training phase.

Table 2: A comparison of the SN and SP values obtained on both training and testing datasets by using various color space components as input to the CNN.

Method	Training data		Testing data	
	SN (in %)	SP (in %)	SN (in %)	SP (in %)
a (CIE-*Lab*)	93	95	25	98
b (CIE-*Lab*)	90	98	16	98
cr (YCbCr)	92	96	20	98
CIE-*Lab*	92	97	21	98

3.2. Optimal Color Channel Selection

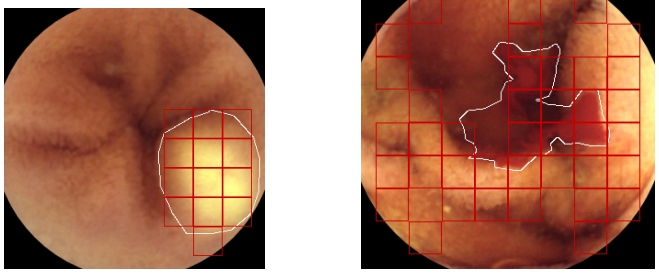
The aim of this experiment is to find the optimal set of color space components that can be used to constitute the input to the CNN. The color spaces CIE-*Lab* and YCbCr, and their components, have been used in the experiment. Patches are extracted from the images and the minority class is expanded using the *reflection* technique. The architecture of the CNN used for this experiment is identical to the one used in Section 3.1. However, for the color space input, a 3D kernel was used in the first convolutional layer. Table 2 presents the results of this experiment. The results for the best three color components and the best color space, according to the evaluation metrics, are presented. It can be observed that the *a* channel of CIE-*Lab* color space gives the best sensitivity for a given specificity, which makes it the color channel of choice for our CNN.

3.3. CNN Architecture and Training

Fig. 2 gives an overview of the optimal CNN architecture chosen for our experiments. The input is normalized as bringing all pixels to the same scale improves convergence. We use three kernels with a planar spread of 5×5 in all convolutional layers and max-pooling layers of size 2×2. We employ a rectified linear unit (ReLU) activation in both convolutional and fully-connected layers except the final layer, which has a sigmoid activation to convert the scores into probabilities for each class. We train the network by minimizing the binary cross-entropy between the predicted label and the true label using the Adam optimizer. Regularization is not employed in the convolutional layers due to their inherent resistance to over-fitting by virtue of shared weights across the entire image. The max-pooling layers, however, are regularized using dropout. Parameters such as the number of convolutional and max-pooling layers, dropout ratio in regularization, etc., have been decided on empirically.

Table 3: A comparison of the AUC (AUC: Area under ROC curve) values calculated for the classification of each disease individually with the current state-of-the-art results.

Disease	Proposed (in %)	Iakovidis et al. (in %)
Aphthae	78.81 ± 10.14	79.1 ± 13.1
Bleeding	64.08 ± 5.39	83.5 ± 10.1
Chylous Cysts	87.85 ± 6.8	87.6 ± 4.3
Lymphangiectasias	95.95 ± 2.28	96.3 ± 3.6
Polypoids	73.86 ± 7.11	85.9 ± 6
Stenoses	76.73 ± 3.65	80.2 ± 13.4
Ulcers	89.4 ± 2.26	76.2 ± 10
Villous Oedema	78.38 ± 7.38	92.3 ± 7.6



(a) Detected abnormality in Chylous Cyst image

(b) Detected abnormality in Bleeding image

Fig. 3: Results of a (a) successful and a (b) failed classification of the proposed method. The red borders mark the patches that have been classified as malignant by our classifier.

4. DISCUSSION

We use the AUC to evaluate the performance of the network over a range of discriminative thresholds. The AUC values are calculated for classification of each disease individually and for holistic classification over all the diseases. We use a 10-fold Monte Carlo cross-validation and compare the results with the current state-of-the-art results (as reported in [8]) in Table 3.

We developed a CNN-based automated classification system, to identify malignant pixels in WCE images, that captures the color and textural information and artificially generates the representative features instead of using hand-crafted features. The overall AUC, SN, and SP for all the diseases are $79.61 \pm 3.5\%$, $71 \pm 19\%$, and $72 \pm 3\%$ respectively. The AUC of the proposed CNN model performs comparably or even outperforms the state-of-the-art method in detecting abnormalities for the classes aphthae, chylous cysts, lymphangiectasias, stenoses, and ulcers, showing that the features of the CNN-based approach are more representative of these abnormalities. However, the performance in detecting abnormalities of the other classes is poorer than the state-of-the-art technique. Fig. 3(a) shows the patches corresponding to a chylous cyst being detected successfully. Fig. 3(b) shows a scenario where the model fails to detect bleeding correctly. A plausible explanation for the under-performance in select abnormalities is the lack of sufficient training data pertaining to those classes, and the lack of prominent texture patterns in the case of bleeding, from which the CNN could learn a distinguishing feature.

Acknowledgements: We are grateful to Prof. Dimitris K. Iakovidis, University of Thessaly, Greece, and Prof. Anastasios Koulaouzidis, The Royal Infirmary of Edinburgh, Endoscopy Unit, for kindly providing us with the WCE dataset and for clarifying several aspects related to their lesion detection algorithm.

5. REFERENCES

[1] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, pp. 417, May 2000.

[2] S. A. Karkanis, D. K. Iakovidis, D. A. Karras, and D. E. Maroulis, "Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial neural network architectures," in *Proceedings of 2001 IEEE International Conference on Image Processing (ICIP)*, vol. 2, pp. 833–836.

[3] G. Lv, G. Yan, and Z. Wang, "Bleeding detection in wireless capsule endoscopy images based on color invariants and spatial pyramids using support vector machines," in *Proceedings of 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6643–6646.

[4] D. K. Iakovidis, D. Chatzis, P. Chrysanthopoulos, and A. Koulaouzidis, "Blood detection in wireless capsule endoscope images based on salient superpixels," in *Proceedings of 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 731–734.

[5] B. Li and M. Q.-H. Meng, "Texture analysis for ulcer detection in capsule endoscopy images," *Image and Vision Computing*, vol. 27, no. 9, pp. 1336–1342, 2009.

[6] B. Li and M. Q.-H. Meng, "Automatic polyp detection for wireless capsule endoscopy images," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10952–10958, 2012.

[7] D. K. Iakovidis, S. Tsevas, and A. Polydorou, "Reduction of capsule endoscopy reading times by unsupervised image mining," *Computerized Medical Imaging and Graphics*, vol. 34, no. 6, pp. 471–478, 2010.

[8] D. K. Iakovidis and A. Koulaouzidis, "Automatic lesion detection in wireless capsule endoscopy- a simple solution for a complex problem," in *Proceedings of 2014 IEEE International Conference on Image Processing (ICIP)*, pp. 2236–2240.

[9] D. K. Iakovidis and A. Koulaouzidis, "Automatic lesion detection in capsule endoscopy based on color saliency: closer to an essential adjunct for reviewing software," *Gastrointestinal Endoscopy*, vol. 80, no. 5, pp. 877–883, 2014.

[10] A. Karargyris and N. Bourbakis, "Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 10, pp. 2777–2786, 2011.

[11] P. Szczypiński, A. Klepaczko, M. Pazurek, and P. Daniel, "Texture and color based image segmentation and pathology detection in capsule endoscopy videos," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 1, pp. 396–411, 2014.

[12] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *arXiv preprint arXiv:1603.06430*, 2016.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.