Bibliographic study focused on pre-processing methods on Wireless Capsule Endoscopy (WCE) images

Tan Sy NGUYEN

February 25, 2021

In WCE image, lesion objects or areas itself can have different shapes, textures, colors and orientations. They can be located anywhere in the frame and also partially be hidden and covered by biological substances, like seeds or stool, and lighted by direct and ambient light. Moreover, the image itself can be interleaved, noisy, blurry and over or under exposed, and it can contain borders and subimages. Apart from that, it can have various resolutions depending on the type of endoscopy equipment used. Endoscopic images usually have a lot of flares and flashes caused by high power light source located close to the camera. All these nuances affect the local features detection methods negatively and have to be specially treated to reduce localisation precision impact.

Fig. 1 illustrates the resulting categorization of research topics in the field of endoscopic image/video pre-processing and analysis, representing the structure of the following sections as well.

1 Image enhancement

A number of publications deal with the enhancement of frames from endoscopic videos in order to improve the visual quality of the video. That means that the underlying data, i.e., the pixels of the individual frames are not only analyzed but also modified while other analysis approaches described in the upcoming sections only try to extract information without changing the content. In this context a number of well-established general purpose image processing techniques can be applied, but this section will focus on techniques and research findings that specifically address the domain of endoscopy. Another aspect that is particularly important in this context is real-time capability because the optimized result should instantly be visible at the screen during a procedure. However, image enhancement and pre-processing is not only interesting for real-time applications but can also be of great importance as a preparation step for any kind of further automatic processing. Early work in this area includes:

- Automatic adjustment of contrast with the help of clustering and histogram modification [1].

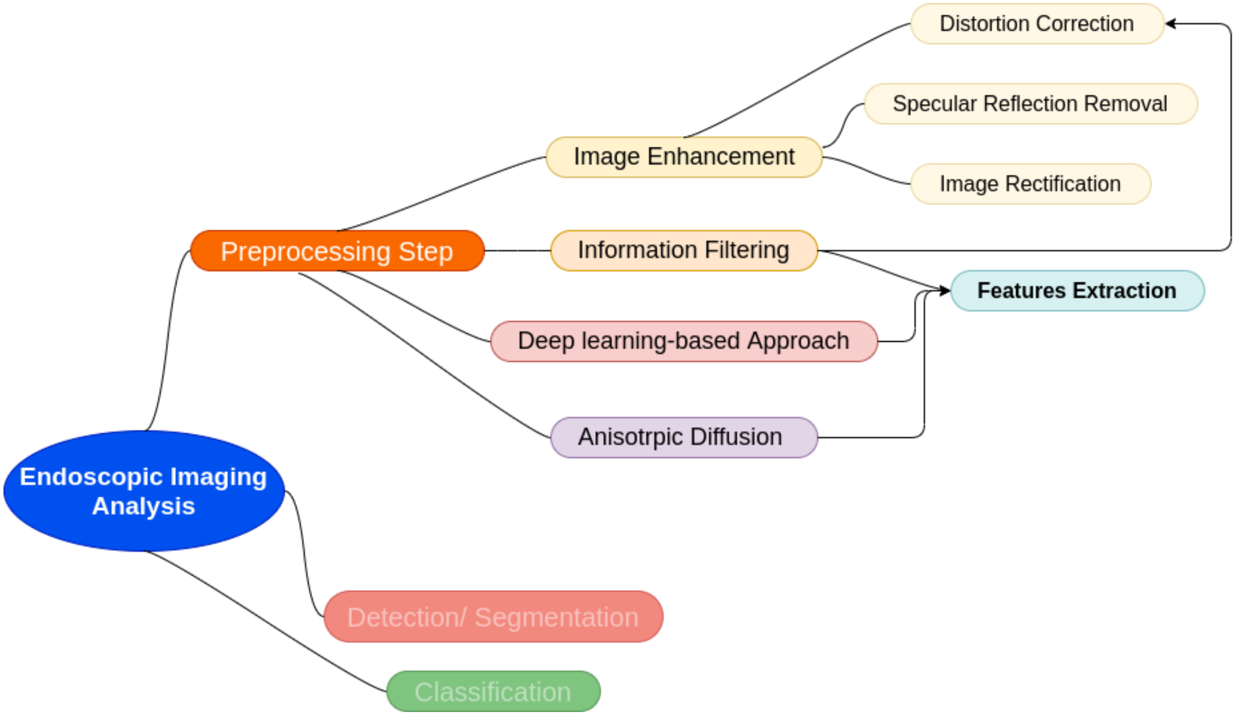
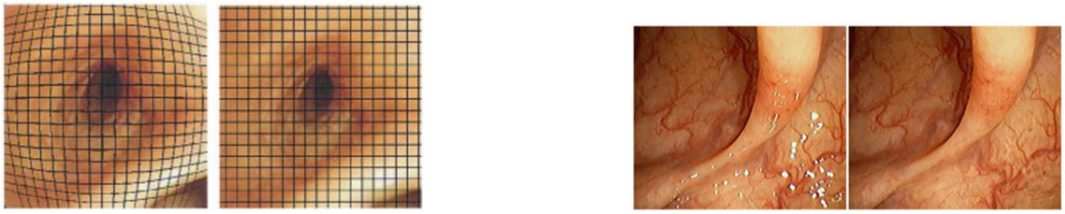


Figure 1: Categorisation of publications in the field of endoscopic video pre-processing

- Removal of temporal noise, i.e., small flying particles or fast moving smoke only appearing for a short moment at one position, by using a temporal median filter of color values [2].
- Color normalization using an affine transformation in order to get rid of a reddish tinge caused by blood during therapeutic interventions and to obtain a more natural color [2]
- Correction of color misalignment: Most endoscopes do not use a color chipset camera but a monochrome chipset that only captures luminance information. To get a color video, red, green and blue color filters have to be applied sequentially. In case of rapid movements - which occur frequently in endoscopic procedures - the color channels become misaligned. This is not only annoying when watching the video but particularly hindering further automatic analysis. Dahyot et al. [3] propose to use color channels equalization, camera motion estimation and motion compensation to correct the misalignments.

1.1 Camera calibration and distortion correction

Typical endoscopes have a fish-eye lens to provide a wide-angle field of view. This characteristic is useful because the endoscopist can see a larger area. However, the drawback is a non-linear geometric distortion (barrel distortion). Objects located in the center of the image appear larger and lines get bended as illustrated in Fig. 2a. This distortion has to be corrected prior to advanced methods that rely on correct geometric information, e.g., 3D



(a) Distortion correction

(b) Specular reflection removal

Figure 2: Illustration of image enhancement methods for endoscopy

reconstruction or image registration. The basic problem is to find the distortion center and the parameters that describe the extent of the distortion, which is not constant but depends on the respective endoscope. This process is also known as camera calibration and includes the determination of intrinsic and extrinsic camera parameters. Vijayan et al. [4] proposed to use a calibration image showing a rectangular grid of dots. This image is captured by the endoscope, resulting in a distorted version of the calibration image. Then the transformation parameters from this distorted image to the original calibration image are calculated using polynomial mapping and least squares estimation. These parameters are used to build a model that can then be used to correct the actual frames from the endoscopic video. This approach was further improved in [5] and [6]. A further approach in [7] is not only applicable to forward viewing endoscopes but also to oblique viewing endoscopes. Their camera model is able to compensate the rotation but has a higher complexity and more parameters. For calibration, they use a chess pattern image instead of a grid of dots. Further publications using this calibration pattern are [8, 9, 10]. In [11], the authors investigate if distortion correction also affects the accuracy of CAD (Computer Aided Diagnosis). The surprising result was that for many feature extraction techniques the performance did not improve but was even worse than without distortion correction. Only for shape-based features that rely on geometrical properties a modest improvement was observed. Further research results in this field can be found in [12, 13, 14].

1.2 Specular reflection removal

Endoscopic images often contain specular light reflections, also called highlights, on the wet tissue surface. They are caused by the inherent frontal illumination and are very distracting for the observer. A study conducted in [2] shows that physicians prefer images where they are corrected. Even worse, reflections severely impair analysis algorithms because they introduce wrong pixel values and additional edges. This also impairs image feature extraction, which is an essential technique for reconstruction, tracking etc. Hence, a number of approaches for correction have been proposed as a supporting component for other analysis methods, e.g., detection of non-informative frames [15], segmentation and detection of surgical instruments [16, 17], tracking of natural landmarks for cardiac motion estimation [18], reconstruction of 3D structures [19] or correction of color channel misalignment [20].

Most approaches consist of two phases. First, the highlights are detected in each frame. This is rather straightforward and in most cases uses basic histogram analysis, thresholding and morphological operations. Pixels with an intensity above a threshold are regarded as highlights. Some authors additionally propose to check for low saturation as a further strong indication for specular highlights ([15, 21]). In this context, the usage of various color spaces has been proposed, e.g., RGB, YUV, HSV, HSI, CIE-xyY. In a second phase, the pixels identified as reflections are "corrected", i.e., modified in a way that the resulting image looks as realistic as possible. An example of a corrected image can be seen in Fig. 2b. An important aspect is that user should be informed about this image enhancement, because one cannot rule out the possibility that wrong information is introduced, e.g., a modified pit pattern on a polyp that can adversely affect the diagnosis.

1.3 Image rectification

In surgical practice, a commonly used type of endoscopes are oblique-viewing endoscopes (e.g., 30°). The advantage of this design is the possibility to easily change the viewing direction by rotating the endoscope around its axis. This enables a larger field of view. The problem is that also the image rotates, resulting in a non-intuitive orientation of the body anatomy. The surgeon has to unrotate the image in their mind in order to not lose their orientation. The missing information about the image orientation is especially a problem in Natural Orifice Translumenal Endoscopic Surgery (NOTES), where a flexible endoscope is used (as opposed to rigid endoscopes like in laparoscopy). Some approaches have been proposed that use modified equipment to tackle this problem, e.g., an inertial sensor mounted on the endoscope tip [22], but hardware modifications always limit the practical applicability. Koppel et al. [23] propose an early vision-based solution. They track 2D image features to estimate the camera motion. Based on this estimation, the image is rectified, i.e., rotated such that the natural "up" direction is restored. Moll et al. [24] improve this approach by using the SURF descriptor (Speeded Up Robust Features), RANSAC (Random Sample Consensus) and a bag-of-visual-words approach based on Integrated Region Matching (IRM). A different approach [25] exploits the fact that endoscopic images often feature a "wedge mark", a small spike outside the image circle that visually indicates the rotation. By detecting the position of this mark, the rotation angle can easily be computed.

2 Information filtering

Endoscopic videos typically contain a considerable amount of frames that do not carry any relevant information and therefore are useless for content-based analysis. Hence, it is desirable to automatically detect such frames and sort them out, i.e., perform a temporal filtering. This can be regarded as a different kind of pre-processing, with the difference that not the pixels of individual frames are modified but the video as such is modified to the effect that frames are removed. This idea is closely related to video summarization which can be seen as an intensification of frame filtering. In video summarization, the goal is to select especially informative frames or sequences and reduce the video to an even higher extent. Moreover, it is often the case that only parts of on image are non-informative, but other regions are indeed relevant for the analysis. To concentrate analysis on such selected regions, several image segmentation techniques have been proposed to perform a spatial filtering.

In the literature, different criteria can be found for a frame to be considered as informative or non-informative. The most important criterion is blurriness. According to [26], about 25 % of the frames of a typical colonoscopy video are blurry. Oh et al. [15] propose to use edge detection and compute the ratio of isolated pixels to connected pixels in the edge image to determine the blurriness. As this method depends very much on the selection of thresholds and further parameters, they propose a second approach using discrete Fourier transformation (DFT). Seven texture features are extracted from the gray-level co-occurrence matrix (GLCM) of the resulting frequency spectrum image and used for k-means clustering to differentiate between blurry and clear images. A similar approach by [26] uses the 2D discrete wavelet transform with a Haar wavelet Kernel to obtain a set of approximation and detail coefficients. The L-2 norm of the detail coefficients of the wavelet decomposition is used as feature vector for a Bayesian classifier. This method is nearly 10-times faster than the DFT-based method and also has a higher accuracy. Rangseekajee and Phongsuphap [27] and Rungseekajee et al. [28] on the other side took up the edge-based approach for the domain of thoracoscopy and added adaptive thresholding as pre-processing step to reduce the effect of lighting conditions. Besides, they claim that the Sobel edge detector is more appropriate for this task than the Canny edge detector because it detects less edges due to irrelevant details caused by noise. Another approach [29] uses inter-frame similarities and the concept of manifold learning for dimensionality reduction to cluster indistinct frames. Grega et al. [30] compared the different approaches for the domain of bronchoscopy and reported results for F-measure, sensitivity, specificity and accuracy of at least 87 % or higher. According to them, the best-scoring alternative is a transformation-based approach using discrete cosine transformation (DCT).

Especially in the context of WCE (Wireless Capsule Endoscopy), the presence of intestinal juices is another criterion for non-informative images. Such images are characterized by bubbles that occlude the visualization field. Vilarino et al. [31] use Gabor filters to detect them. According to their studies, 23 % of all images can be discarded, meaning that the visualization time for the manual diagnostic assessment as well as the processing time for automatic diagnostic support can be considerably reduced. In [32], a similar approach is proposed that uses a Gauss Laguerre transform (GLT)-based multiresolution texture feature and introduces a second step that uses spatial segmentation of the bubble region to classify ambiguous frames.

A further type of non-informative frames are out-of-patient frames, i.e., frames from scenes that are recorded outside the patients body. They often occur at the beginning or end of a procedure because it is not always possible to start and stop the recording exactly at the right time. The need for manual recording triggering in general deters many endoscopists from recording videos at all. To address this issue, [33] propose a system that automatically detects when a colonoscopic examination begins and ends. Every time a new procedure is detected, the system starts recording and writes a video file to the disk until the end of the procedure is detected. The proposed approach uses simple color features that work well for the domain of colonoscopy. In [34], the authors extend their approach by various temporal features that take into account the amount of motion to avoid false positives.

To synthesis some results developed depending on preprocessing method, I chose one paper to present and display some detection-based papers results in the table below.

Examples: Automatic Classification Based on Features Fusion for Upper Gastrointestinal WCE Images [41].

Publ./System	Abnormality	pre-processing type	Recall	Precision	Specificity	Accuracy
Wang et al. [35]	polyp	Information filter	81.4%	-	-	-
Mamonov et al. [36]	polyp	Normalization of intensity	47%	-	90%	-
Hwang et al. [34]	polyp	Information filter	96%	83%	-	-
Li and Meng [37]	Tumor	Wavelet transform	88.6%	-	96.2%	92.4%
Zhou et al. [38]	polyp	RGB averaging	75%	-	95.92%	90.77%
Ameling et al. [39]	polyp	Information filter	AUC = 95%	-	-	-
Michael et al. [40]	polyp	Information filter	98.5%	93.88%	72.49%	87.70%
Min et al. [41]	5 types [41]	Label Shadow and Highlight	99.01%	99.00%	-	98.99%

Table 1: A performance comparison of detection approaches. Not all performance measurements are available for all methods, but including all available information gives an idea about each preprocessing step

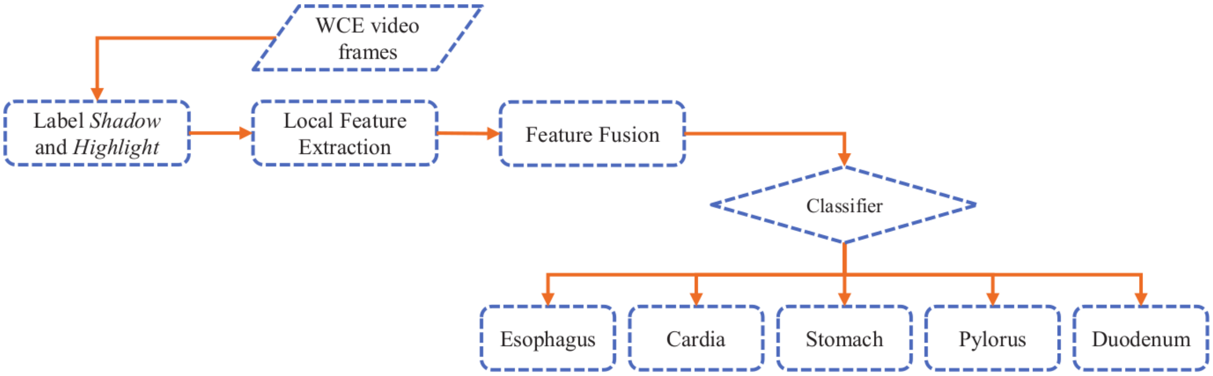


Figure 3: Automatic Classification Based on Features Fusion for Upper Gastrointestinal WCE Images pipeline.

In this paper, a scheme for feature fusion is developed, which improves classification accuracy by using a statistical-based analysis approach discrimination power analysis (DPA)[42]. In addition, WCE images suffer from illumination variations due to the non-ambient lighting, and thus, a new colour scale invariant local ternary pattern (CSILTP) algorithm is proposed to extract features of WCE images. CSILTP is an extension of SILTP[43] which is more robust to illumination variation. The main contributions can be summarized as follows.

- **An automatic tuning shadow and highlight detection algorithm based on superpixel-level is proposed.**
- A new feature descriptor CSILTP is extracted as a texture feature to improve the accuracy of WCE images classification.
- An efficient scheme is developed to fuse multiple local features by employing DPA.

Label Shadow and Highlight

Since the tubular cavity structure of the GI tract, the lumen is expected to appear darker than surroundings in view. And the tissues closer to the light source are imaged with extreme bright due to the non-ambient lighting. Those regions are refer to as shadow and highlight, respectively [44]. The results of organ classification are unreliable if the whole regions in

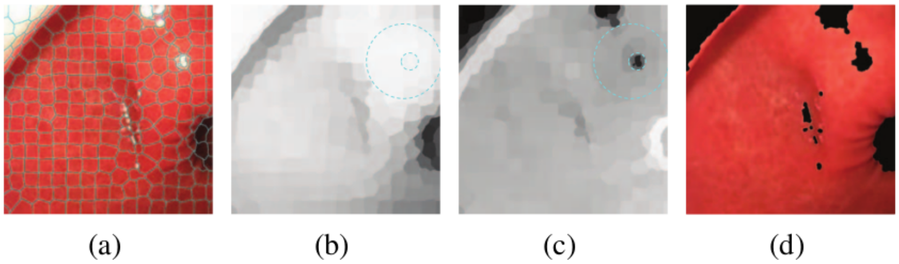


Figure 4: Illustration of detection of shadow and highlight. (a) Original image with 300 superpixels. (b) Average intensity superpixel map of original image. (c) Average saturation superpixel map of original image. (d) Image labeled valid regions (original is segmented by 1500 superpixels and the invalid regions are shown in black)

images are used to extract features [45]. Thus, the regions of shadow and highlight should be labeled and removed before feature extraction.

A novel automatic tuning shadow and highlight detection algorithm is proposed in this paper, which based on superpixel segmentation. Simple linear iterative clustering (SLIC) algorithm [46] is first utilized to obtain superpixels, as shown in Fig. 4(a). Every superpixel's intensity is assigned a value which equals to the average of all pixels in the surperpixel. Then, the color image is split into three intensity (gray scale) images (i.e. hue, saturation and intensity in HSI model). As shown in Fig. 4(b), the intensities of the shadow are relatively small (shown in dark gray). Thus, the shadow are where the intensities are less than a certain value (i.e. the threshold). However, threshold selection is a paradoxical process since a fixed threshold cannot be applied to all images. Thus, an automatic tuning threshold algorithm is proposed.

3 Anisotropic diffusion based method

Anisotropic contrast diffusion is employed to contrast the images due to low dark quality of the WCE images. It also able to make characteristics in a WCE image is more visible by human eyes also by computer machine. Contrast enhancement method is employed to make more contrast on the WCE images compared to these using the original concept introduced by Li et al. [47]. In [48], Shahril et al. introduce a variance formula to overcome the B. Li's weaknesses. In order to make the details of each image to be more visible, sharpening algorithm is proposed to ease the classification process of the abnormalities such as bleeding in WCE images. The following table showing us the comparison results of the given methods. Note that we are focusing on pre-processing stage for the real life situation. I would like to choose one method as example to explain us what they have done to enhance the image. **Example**: Wireless capsule endoscopy images enhancement via adaptive contrast diffusion [47]. In order to get a contrast description of one point in an image, they resort to local analysis of an image by using Hessian matrix. Hessian matrix of one point in a gray image under a given scale σ is:

$$\mathbf{H}_\sigma(x,y) = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{bmatrix} \tag{1}$$

Publ./System	Abnormality	PSNR	Recall	Precision	Specificity	Accuracy
Li et al. [47]	4 patients [47]	22.57	69.5%	-	73.5%	71.5%
Shahril et al. [48]	4 patients [47]	22.47	-	-	-	-

Table 2: A performance comparison of detection approaches. Not all performance measurements are available for all methods, but including all available information gives an idea about each preprocessing step

where I_{xx}, I_{xy}, I_{yy} are the second-order derivative of the image along direction of x, y, xy respectively. Here σ is implicitly contained in the calculation of second-order derivatives. Assume that Hessian matrix of one point has two eigenvalues: $\lambda_1(x, y)$ and $\lambda_2(x, y)$. Considering the fact that intensity variation in background of an image is rather weak, they may conclude that differential values of such regions are small, which result in small eigenvalues. On the other hand, in regions where there is rather apparent intensity variations compared with background, they may draw an opposite conclusion.

A new concept of contrast is established as follows:

$$c(x, y) = \lambda_1^2(x, y) + \lambda_2^2(x, y) \tag{2}$$

It characterizes intensity variations from the standpoint of Hessian matrix eigenvalues. Applying this concept to the whole image, the image's corresponding contrast space is taken. Employing this contrast space, anisotropic diffusion is changed into:

$$\frac{\delta c(x, y, t)}{\delta} = div[g(c)\nabla_c] = g\nabla_c + \nabla_g.\nabla_c \tag{3}$$

where $g(c)$:

$$g(c) = \frac{1}{1 + (\|c\|/k)^2} \tag{4}$$

where k is the contrast parameters. It determines behavior of the diffusion according to the value of contrast in the region. After diffusing in the contrast space, the diffused result is transformed back to image space by normalization as illustrated below:

$$I = \frac{c - c_{min}}{c_{max} - c_{min}} \times 255 \tag{5}$$

4 Deep learning based methods

It is clear to see the current bad situation of WCE image quality. The quality of image data is often reduced due to overlays of text and positional data. In [49], Kirkerod et al. present different methods of preprocessing such images and they describe their approach to GI disease classification for the Kvasir v2 dataset. Multiple approaches are proposed to inpaint problematic areas in the images to improve the anomaly classification, and they discuss the effect that such preprocessing does to the input data.

Neural networks, in the context of deep learning, show much promise in becoming an important tool with the purpose assisting medical doctors in disease detection during patient examinations. However, the current state of deep learning is something of a ”black box”, making it very difficult to understand what internal processes lead to a given result. This is not only true for non-technical users but among experts as well. This lack of understanding has led to hesitation in the implementation of these methods among mission-critical fields, with many putting interpretability in front of actual performance. Motivated by increasing the acceptance and trust of these methods, and to make qualified decisions, we present a system that allows for the partial opening of this black box. This includes an investigation on what the neural network sees when making a prediction, to both, improve algorithmic understanding, and to gain intuition into what pre-processing steps may lead to better image classification performance. Furthermore, a significant part of a medical expert’s time is spent preparing reports after medical examinations, and if we already have a system for dissecting the analysis done by the network, the same tool can be used for automatic examination documentation through content suggestions. To deal with this challenge, Hicks et al. [50] present a system that can look into the layers of a deep neural network and present the network’s decision in a way that that medical doctors may understand. Furthermore, we present and discuss how this information can possibly be used for automatic reporting.

Publ./System	**Abnormality**	**pre-processing type**	**Recall**	**Precision**	**Specificity**	**Accuracy**
Min et al. [49]	Kvasir dataset v1 [51]	Auto-generate by autoencoder	93.94%	93.96%	-	-
Hicks et al. [50]	Kvasir dataset v2 [51]	Layer by layer visialization	94.30%	96.80%	74.90%	94.30%

Table 3: A performance comparison of detection approaches. Not all performance measurements are available for all methods, but including all available information gives an idea about each preprocessing step

I chose one example to explain one detail what they have done in pre-processing step.

Example: Dissecting Deep Neural Networks for Better Medical Image Classification and Classification Understanding [50].

The analysis performed by the system is based on deep learning technologies, specifically CNNs. These are used to analyse image or video data to perform different classification tasks, e.g., automatic detection of diseases. This process is made transparent to the users through the neural network dissection tool, which examines the individual layers of a CNN and allows for inspection of what regions of an image contribute to the score of a given class. This transparency is a critical piece in building trust and acceptability among non-technical users of the system, like medical experts, who rely on the system’s output without detailed knowledge of the underlying processes. Furthermore, it allows for discovering faults within the trained model and the dataset used to train the system.

Visualisations are primarily based on the weighted gradient class activation map (grad-CAM) technique [52], which allows for visualisation of different CNN architectures without the need for modifications (replacement of layers). An overview of this process is shown in Fig. 5, and starts once the user has selected an input image, target layer and target class using the web-interface. Based on these parameters, the system generates three different representations of the given image (all shown in Fig. 6)

The visualisation process starts once the user has selected an image, layer, and class for further analysis. With this set, they begin with the creation of the grad-CAM and saliency visualisations. Starting with the grad-CAM, they calculate the gradient of the target layer

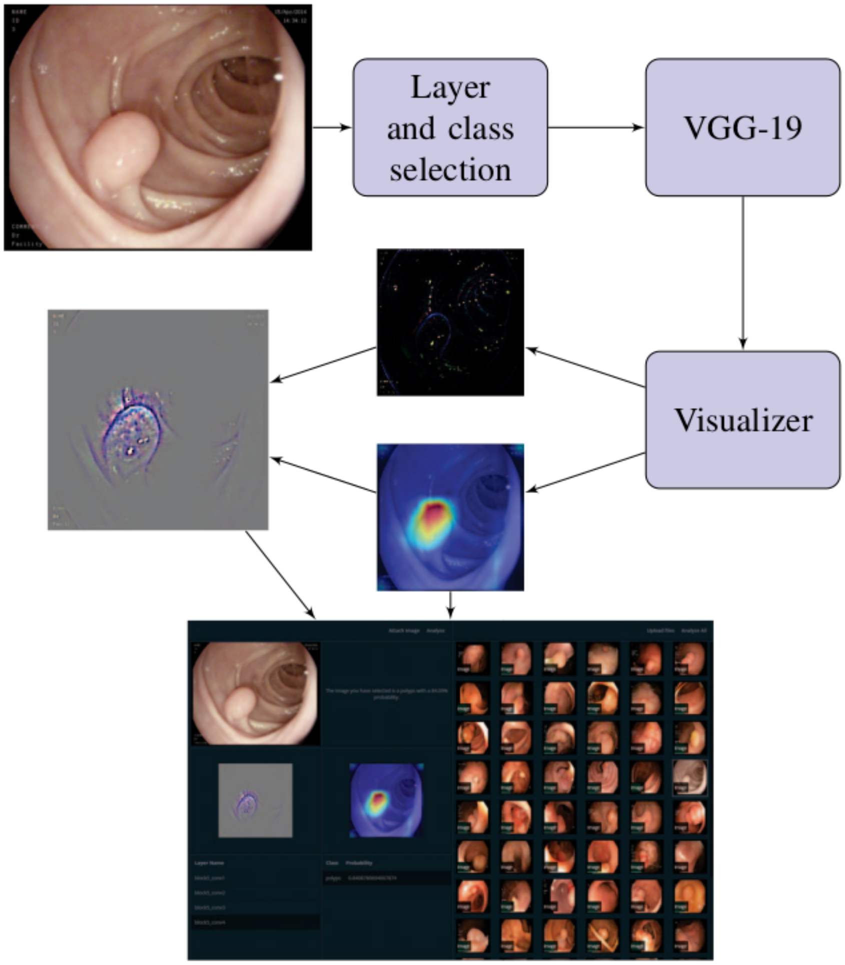
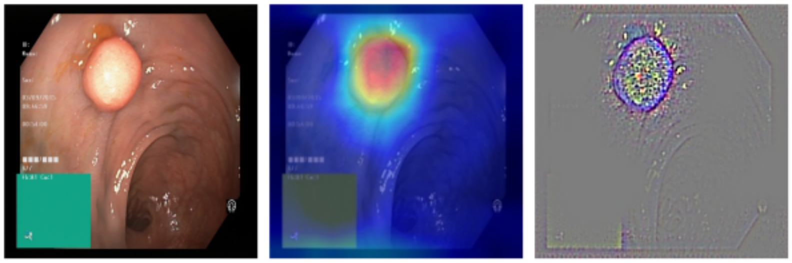


Figure 5: An overview of how they produce the two visualisations included in the image analysis, and how it is presented in the user interface where a visualisation of the different convolutional layers can be selected.



(a) Original (b) Grad-CAM (c) Grad-CAM

Figure 6: n image that has been correctly identified as containing a polyp by our CNN, together with the grad-CAM and guided grad-CAM representation.

using the loss of the selected class in regards to the input image. These gradients are then globally average pooled to get the weights, which is then multiplied by the output of the target layer and passed through a ReLU function to produce the grad-CAM representation. The grad-CAM is then re-sized back to the dimensions of the original image and has its values squashed between 0 and 1 before applying a blue-red heat map.

To generate the guided back-propagation saliency map, they replace the activations of the original network with a modified ReLU function. During back-propagation, a conventional ReLU would let all gradients whose inputs were larger than 0 pass. This rule is extended by additionally discarding all gradients that are below 0, thereby only back-propagating the positive influence on the activations. With this modified network, we calculate the gradients of the target layer with respect to the input image, i.e., these gradients represent our saliency map. With the grad-CAM and saliency map generated, we multiply them together to produce the guided grad-CAM representation. This together with the grad-CAM is used in proposed system.

References

[1] S. Sheraizin and V. Sheraizin. “Endoscopy Imaging Intelligent Contrast Improvement”. In: (2005), pp. 6551–6554.

[2] Florian Vogt et al. “A System for Real-Time Endoscopic Image Enhancement”. In: (2003). Ed. by Randy E. Ellis and Terry M. Peters, pp. 356–363.

[3] Rozenn Dahyot, Fernando Vilariño, and Gerard Lacey. “Improving the Quality of Color Colonoscopy Videos”. In: *EURASIP J. Image and Video Processing* 2008 (Jan. 2008). DOI: 10.1155/2008/139429.

[4] K. Vijayan Asari, S. Kumar, and D. Radhakrishnan. “A new approach for nonlinear distortion correction in endoscopic images based on least squares estimation”. In: *IEEE Transactions on Medical Imaging* 18.4 (1999), pp. 345–354.

[5] Chao Zhang et al. “Nonlinear distortion correction in endoscopic video images”. In: 2 (2000), 439–442 vol.2.

[6] J. P. Helferty et al. “Videoendoscopic distortion correction and its application to virtual guidance of endoscopy”. In: *IEEE Transactions on Medical Imaging* 20.7 (2001), pp. 605–617.

[7] Tetsuzo Yamaguchi et al. “Camera Model and Calibration Procedure for Oblique-Viewing Endoscope”. In: (2003). Ed. by Randy E. Ellis and Terry M. Peters, pp. 373–381.

[8] Thomas Stehle et al. “Dynamic Distortion Correction for Endoscopy Systems with Exchangeable Optics”. In: (Jan. 2009). DOI: 10.1007/978-3-540-93860-6_29.

[9] Joao P Barreto et al. “Automatic Camera Calibration Applied to Medical Endoscopy”. In: *British Machine Vision Conference, BMVC 2009 - Proceedings* (Sept. 2009). DOI: 10.5244/C.23.52.

[10] J. P. Barreto, R. Swaminathan, and J. Roquette. “Non Parametric Distortion Correction in Endoscopic Medical Images”. In: (2007), pp. 1–4.

[11] M. Gschwandtner et al. “Experimental study on the impact of endoscope distortion correction on computer-assisted celiac disease diagnosis”. In: (2010), pp. 1–6.

[12] Nicole Kallemeyn et al. “Arthroscopic Lens Distortion Correction Applied to Dynamic Cartilage Loading”. In: *The Iowa orthopaedic journal* 27 (Feb. 2007), pp. 52–7.

[13] M. Liedlgruber, A. Uhl, and A. Vécsei. “Statistical analysis of the impact of distortion (correction) on an automated classification of celiac disease”. In: *2011 17th International Conference on Digital Signal Processing (DSP)* (2011), pp. 1–6. DOI: 10.1109/ICDSP.2011.6004900.

[14] Christian Wengert et al. “Fully Automatic Endoscope Calibration for Intraoperative Use”. In: *Bildverarbeitung fur Die Medizin* (Jan. 2006), pp. 419–423. DOI: 10.1007/3-540-32137-3_85.

[15] JungHwan Oh et al. “Informative frame classification for endoscopy video”. In: *Medical Image Analysis* 11.2 (2007), pp. 110–127. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2006.10.003. URL: http://www.sciencedirect.com/science/article/pii/S136184150600079X.

[16] Charles-Auguste Saint-Pierre et al. “Detection and Correction of Specular Reflections for Automatic Surgical Tool Segmentation in Thoracoscopic Images”. In: *Mach. Vision Appl.* 22.1 (Jan. 2011), pp. 171–180. ISSN: 0932-8092.

[17] Y. Cao et al. “Computer-Aided Detection of Diagnostic and Therapeutic Operations in Colonoscopy Videos”. In: *IEEE Transactions on Biomedical Engineering* 54.7 (2007), pp. 1268–1279.

[18] Martin Gröger, Tobias Ortmaier, and Gerd Hirzinger. “Structure Tensor Based Substitution of Specular Reflections for Improved Heart Surface Tracking”. In: (2005). Ed. by Hans-Peter Meinzer et al., pp. 242–246.

[19] D. Stoyanov and Guang Zhong Yang. “Removing specular reflection components for robotic assisted laparoscopic surgery”. In: 3 (2005), pp. III–632.

[20] Arnold Mirko et al. “Automatic Segmentation and Inpainting of Specular Highlights for Endoscopic Imaging”. In: *EURASIP Journal on Image and Video Processing* 2010 (Jan. 2010). DOI: 10.1155/2010/814319.

[21] R. Yao et al. “Specular Reflection Detection on Gastroscopic Images”. In: (2010), pp. 1–4.

[22] K. Holler et al. “Clinical evaluation of Endorientation: Gravity related rectification for endoscopic images”. In: (2009), pp. 713–717.

[23] D. Koppel, Yuan-Fang Wang, and Hua Lee. “Image-based rendering and modeling in video-endoscopy”. In: (2004), 269–272 Vol. 1.

[24] Markus Moll et al. “Unrotating images in laparoscopy with an application for 30° laparoscopes”. In: 22 (Jan. 2009), pp. 966–969. DOI: 10.1007/978-3-540-89208-3_230.

[25] N. Fukuda et al. “A scope cylinder rotation tracking method for oblique-viewing endoscopes without attached sensing device”. In: (2010), pp. 684–687.

[26] M. Arnold et al. "Indistinct Frame Detection in Colonoscopy Videos". In: *2009 13th International Machine Vision and Image Processing Conference* (2009), pp. 47–52. DOI: 10.1109/IMVIP.2009.16.

[27] Nicharee Rangseekajee and Sukanya Phongsuphap. "Endoscopy video frame classification using edge-based information analysis". In: 38 (Jan. 2011).

[28] N. Rungseekajee, M. Lohvithee, and I. Nilkhamhang. "Informative frame classification method for real-time analysis of colonoscopy video". In: *2009 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* 02 (2009), pp. 1076–1079. DOI: 10.1109/ECTICON.2009.5137231.

[29] Selen Atasoy et al. "Endoscopic Video Manifolds". In: *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 13 (Sept. 2010), pp. 437–45. DOI: 10.1007/978-3-642-15745-5_54.

[30] Michal Grega et al. "Algorithms for Automatic Recognition of Non-informative Frames in Video Recordings of Bronchoscopic Procedures". In: *Advances in Intelligent and Soft Computing* 69 (Jan. 2010). DOI: 10.1007/978-3-642-13105-9_53.

[31] F. de Iorio et al. "Automatic Detection of Intestinal Juices in Wireless Capsule Video Endoscopy". In: *18th International Conference on Pattern Recognition (ICPR'06)* 4 (2006), pp. 719–722. DOI: 10.1109/ICPR.2006.296.

[32] Md. Khayrul Bashar et al. "Automatic detection of informative frames from wireless capsule endoscopy images". In: *Medical image analysis* 14 (June 2010), pp. 449–70. DOI: 10.1016/j.media.2009.12.001.

[33] Sean Stanek et al. "Automatic real-time capture and segmentation of endoscopy video". In: *Proceedings of SPIE - The International Society for Optical Engineering* (Mar. 2008). DOI: 10.1117/12.770930.

[34] Sean Stanek et al. "Automatic real-time detection of endoscopic procedures using temporal features". In: *Computer methods and programs in biomedicine* 108 (May 2011), pp. 524–35. DOI: 10.1016/j.cmpb.2011.04.003.

[35] Y. Wang et al. "Part-Based Multiderivative Edge Cross-Sectional Profiles for Polyp Detection in Colonoscopy". In: *IEEE Journal of Biomedical and Health Informatics* 18.4 (2014), pp. 1379–1389. DOI: 10.1109/JBHI.2013.2285230.

[36] A. V. Mamonov et al. "Automated Polyp Detection in Colon Capsule Endoscopy". In: *IEEE Transactions on Medical Imaging* 33.7 (2014), pp. 1488–1502. DOI: 10.1109/TMI.2014.2314959.

[37] B. Li and M. Q. -. Meng. "Tumor Recognition in Wireless Capsule Endoscopy Images Using Textural Features and SVM-Based Feature Selection". In: *IEEE Transactions on Information Technology in Biomedicine* 16.3 (2012), pp. 323–329. DOI: 10.1109/TITB.2012.2185807.

[38] M. Zhou et al. "Polyp detection and radius measurement in small intestine using video capsule endoscopy". In: *2014 7th International Conference on Biomedical Engineering and Informatics* (2014), pp. 237–241. DOI: 10.1109/BMEI.2014.7002777.

[39] Stefan Ameling et al. “Texture-Based Polyp Detection in Colonoscopy”. In: *Informatik aktuell* (Jan. 2009), pp. 346–350. DOI: 10.1007/978-3-540-93860-6_70.

[40] M. Riegler et al. “EIR — Efficient computer aided diagnosis framework for gastrointestinal endoscopies”. In: *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)* (2016), pp. 1–6. DOI: 10.1109/CBMI.2016.7500257.

[41] M. Yu et al. “Automatic Classification Based on Features Fusion for Upper Gastrointestinal WCE Images”. In: *2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS)* (2019), pp. 510–515. DOI: 10.1109/DDCLS.2019.8908920.

[42] Saeed Dabbaghchian, Masoumeh Ghaemmaghami, and Ali Aghagolzadeh. “Feature extraction using discrete cosine transform and discrimination power analysis with a face recognition technology”. In: *Pattern Recognition* 43 (Apr. 2010), pp. 1431–1440. DOI: 10.1016/j.patcog.2009.11.001.

[43] Shengcai Liao et al. “Modeling Pixel Process with Scale Invariant Local Patterns for Background Subtraction in Complex Scenes”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 0 (June 2010), pp. 1301–1306. DOI: 10.1109/CVPR.2010.5539817.

[44] X. Zabulis, A. A. Argyros, and D. P. Tsakiris. “Lumen detection for capsule endoscopy”. In: *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2008), pp. 3921–3926. DOI: 10.1109/IROS.2008.4650969.

[45] R. Zhou et al. “Wireless capsule endoscopy video automatic segmentation”. In: *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)* (2012), pp. 825–830. DOI: 10.1109/ROBIO.2012.6491070.

[46] R. Achanta et al. “SLIC Superpixels Compared to State-of-the-Art Superpixel Methods”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pp. 2274–2282. DOI: 10.1109/TPAMI.2012.120.

[47] Baopu Li and Max Meng. “Wireless capsule endoscopy images enhancement via adaptive contrast diffusion”. In: *J. Visual Communication and Image Representation* 23 (Jan. 2012), pp. 222–228. DOI: 10.1016/j.jvcir.2011.10.002.

[48] Rosdiana Shahril, Sabariah Baharun, and A.K.M. Islam. “Pre-processing Technique for Wireless Capsule Endoscopy Image Enhancement”. In: *International Journal of Electrical and Computer Engineering (IJECE)* 6 (Aug. 2016), pp. 1617–1626. DOI: 10.11591/ijece.v6i4.9688.

[49] M. Kirkerod et al. “Unsupervised preprocessing to improve generalisation for medical image classification”. In: *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)* (2019), pp. 1–6. DOI: 10.1109/ISMICT.2019.8743979.

[50] S. Hicks et al. “Dissecting Deep Neural Networks for Better Medical Image Classification and Classification Understanding”. In: *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)* (2018), pp. 363–368. DOI: 10.1109/CBMS.2018.00070.

[51] Konstantin Pogorelov et al. “KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection”. In: *Proceedings of the 8th ACM on Multimedia Systems Conference.* MMSys’17 (2017), pp. 164–169. DOI: 10.1145/3083187.3083212. URL: http://doi.acm.org/10.1145/3083187.3083212.

[52] R. R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 618–626. DOI: 10.1109/ICCV.2017.74.