Artefact Detection using Deep Learning In Endoscopy

Nikhil Agrawal (17117050)
nagrawal@me.iitr.ac.in

Gaurav Singhal (17117027)
gsinghal@me.iitr.ac.in

Kshitij Sharma (17117043)
ksharma@me.iitr.ac.in

Jatin Sharma (17117036)
jsharma@me.iitr.ac.in

Prof. Ankit Bansal (Supervisor)
abansfme@iitr.ac.in

June 6, 2020

Abstract

Endoscopy is a common non-invasive procedure used for both diagnosis and various minimal surgical procedures. The process involves a probe with a camera at the front which is inserted in the targeted body cavity. The procedure may require clear static-camera images or even live video feed but, the camera feed gets routinely distorted. It is distorted by both physical factors like bubbles and debris along with software defects like pixel saturation and motion blur. Since these involve environmental factors, they cannot be merely solved by upgrading the hardware. This leaves room for the application of deep learning techniques, which give a probable look of the body cavity in the absence of defects. We propose a fully automatic framework that can: 1) detect and classify seven different primary defects, 2) provide a quality score for each frame. Existing state-of-the-art methods only deal with the detection of very domain-specific images. In an attempt to find the best method for our use-case we employ 3 different deep learning architectures and modify them according to our instance. In the end, we do a detailed analysis of these methods along with their pros and cons.

Keywords: *Endoscopy; Deep Neural Networks; Object Detection*

1 Introduction

Endoscopy is a nonsurgical procedure used to examine a person's digestive tract. Using an endoscope, a flexible tube with a light and camera attached to it, the doctor can view pictures of your digestive tract on a color TV monitor. The procedure aids in the detection of several gastric abnormalities such as ulcers, inflammation, tumors, and possibly cancerous conditions. In addition, the doctor can perform biopsy along with endoscopy for tissue removal. The aforementioned facts highlight the fact that the success rate is highly dependent on the image quality received by the doctor, which are in most cases corrupted by the presence of numerous artifacts (motion blur, defocus blur, debris, bubbles). Many hindrances caused by these artifacts pose a significant barrier for the doctor to overcome and cause problems even for the postoperative diagnosis. While the highly skilled and experienced doctors can somewhat manage, the rest face a problem as these artifacts even inhibit the use of computer-aided tools hence preventing the use of endoscopy on a large scale.

The objective of our model[1] is the detection and rectification of the above-mentioned image corrupting artifacts posing a problem for endoscopic procedures. The model envisioned would provide real-time clear and corrected images from the endoscope and will be available for the doctor (experienced or not) greatly increases the chances of a successful operation. The accurate detection of artifacts in clinical endoscopy is a critical breakthrough that will transform and accelerate the development of effective endoscopic analysis across all diseases, organs, and modalities. We believe the availability of clear and artifact-free videos for the operation could be the first step towards the automation for endoscopic procedures which could be of great assistance to the doctor.

The existing models for endoscopic workflow have one common limitation - they all address a single artifact class, but naturally, each image has multiple classes present simultaneously. We have used several state of the art models like Yolov3 [3], RetinaNet [2], and Faster RCNN [4] for the detection of multiple artifacts.

The remainder of this article is organized as follows. In **section 2** we introduce our endoscopy data set for artifact detection while **section 3** details our proposed approaches for artifact detection. In **section 4**, we detail the metrics used to evaluate our framework. In **section 5** we present experiments and results for each step of our framework to show the efficacy of individual methods. And finally in **section 6** we conclude the paper and outline directions for future work.

2 Dataset

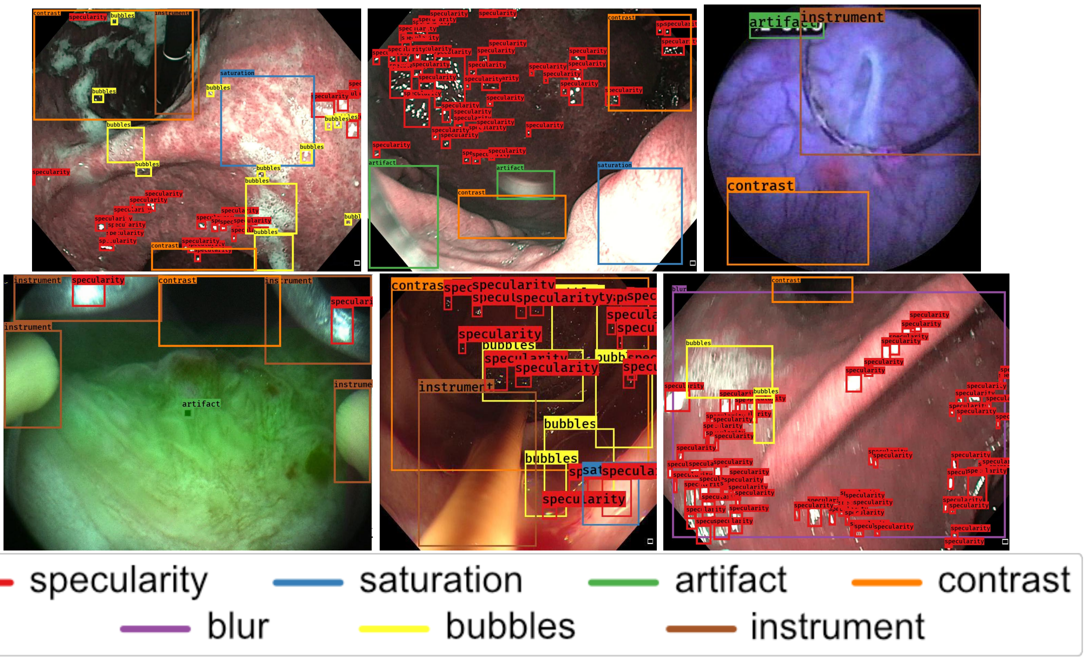


Figure 1: Example annotated training detection boxes illustrating the 8 different artifact classes.

[1] The code is available at https://github.com/nikhilagrawal2000/Artefact-Detection-in-Endoscopy-Images

Our artefact detection data set [1] consists of a total of 2200 endoscopy images. The selection was based on a number of representative artefacts present in these videos and the texture variability of the underlying oesophagus. Two experts annotated a total of 7316 artefacts using bounding boxes where each annotation is classified as:

1. **blur** - streaking from fast camera motion
2. **bubbles** - water bubbles that distort the appearance of the underlying tissue
3. **specularity** - mirror-like surface reflection
4. **saturation** - overexposed bright pixel areas
5. **contrast** - low contrast areas from underexposure or occlusion
6. **blood** - blood being present on the tissues/screen
7. **instruments** - instrument parts which are visible in the image
8. **misc. artefact** (also referred to as 'artefact ' in this paper)- miscellaneous artefacts; e.g., chromatic aberration, debris, imaging artefacts etc.

A 90%-10% split was used to construct the train-validation set for object detection. In general, the training and testing data exhibit the same class distribution and similar bounding boxes (roughly square) but either small with average widths less than 0.2 or large with widths greater than 0.5. Multiple annotations are used in case a given region contains multiple artefacts.

3 Methodology

3.1 Overall Approach

The step-by-step procedure for automatic detection of multiple artefacts and frame restoration of endoscopic videos is presented in Fig 2. It is to be noted that a single frame can be corrupted by multiple artefacts and each artefact class can affect endoscopic frames differently. Therefore, their restoration process is very likely to affect the final restoration results.

Due to less amount of training data, we used various data augmentation techniques like horizontal flip, rotation, scaling, shearing, translation. The corresponding bounding boxes were calculated as per the data augmentation technique.

3.2 Artefact Region Detection

Recent research in computer vision provides us with object detectors that are both robust and suitable for real-time applications. Here, we propose to use a multi-scale deep object detection model for identifying the different artefacts.

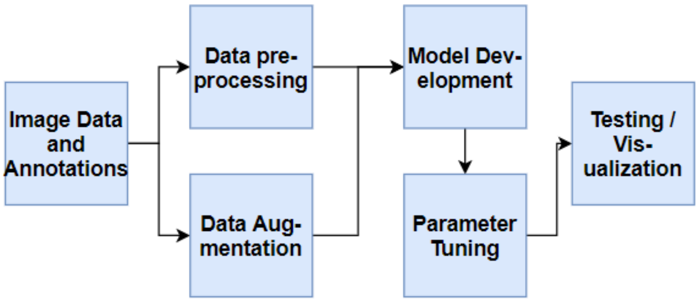


Figure 2: This figure shows the end-to-end framework used in our research.

Faster R-CNNs [4] first introduced a fully trainable end-to-end network yielding an initial region proposal network and successive classifications of the proposed regions without intermediate processing. Since region proposal generation precedes bounding box detection sequentially, this architecture is known as a two-stage detector. Though very accurate, a primary drawback is its slow inference and extensive training. You Only Look Once (YOLO) [3] simplified Faster R-CNNs to predict simultaneously class and bounding box coordinates using a single CNN and a single loss function with good performance and significantly faster inference time. This simultaneous detection is known as a one-stage detector. Compared to two-stage detectors, single-stage detectors mainly suffer two issues: high false detection due to 1) presence of varied size objects and 2) high initial number of anchor boxes requirement that necessitates more accurate positive box mining. The former is corrected by predicting bounding boxes at multiple scales using feature pyramids. To address the latter, RetinaNet [2] introduced a new focal loss which adjusts the propagated loss to focus more on hard, misclassified samples. Recently, YOLOv3 simplified the RetinaNet architecture with further speed improvements. Bounding boxes are predicted only at 3 different scales (unlike 5 in RetinaNet) utilizing objectness score and an independent logistic regression to enable the detection of objects belonging to multiple classes unlike focal loss in RetinaNet. Collectively, Faster R-CNN, RetinaNet and YOLOv3 define the current state-of-the-art detection envelope of accuracy vs speed on the popular natural images benchmark on the COCO Dataset.

We investigated the Faster R-CNN, RetinaNet and YOLOv3 architectures for artefact detection. Validated open source codes are available for all of these architectures.

4 Evaluation

4.1 Quality Assessment metrics

To evaluate our artefact detection we use the standard mean average precision (mAP) and regression loss. Since, time is also a major factor in deciding whether the current architecture can be used in real time we evaluate our model on the basis of mean time taken per epoch.

1. **mAP** - It measures how accurate the prediction is i.e. the percentage of correct predictions.

2. **Regression Loss** - It can be understood as the difference of area between the actual bounding box and predicted bounding box.

3. **Mean Time** - Mean time taken to complete a single epoch.

5 Experiments and Results

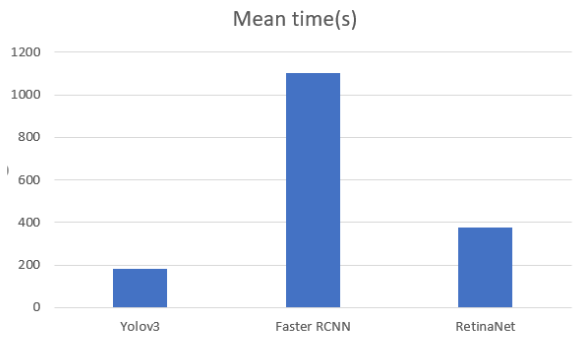


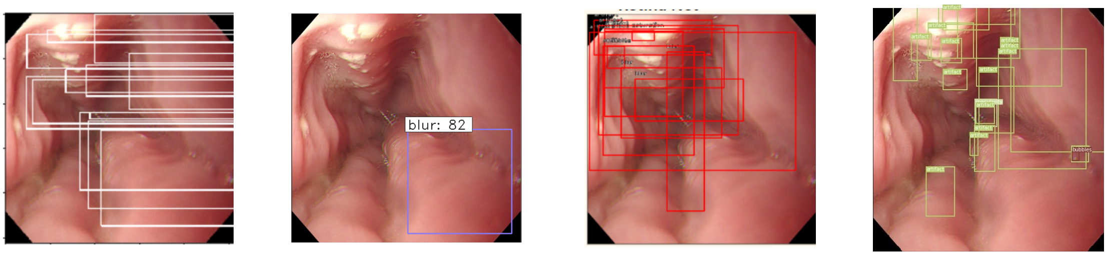
Figure 3: Mean Time of 1 epoch for different architectures

To evaluate the artifact detection, we use the standard mean average precision (mAP). We quantitatively compare the object classification results of all architectures using the mAP and the bounding box detection using regression loss. We have also compared the inference time for object detection in one image.

Table 1: Artifact detection results on validation set with different neural network architectures

Architecture	**mAP**	**Regression Loss**	**Mean Time**
Faster RCNN [4]	0.36	0.1772	1104s
YOLOv3 [3]	0.43	0.5132	182s
RetinaNet [2]	0.24	0.3079	374s

As seen in the (Figure 3) YOLOv3 outperforms both Faster RCNN and RetinaNet in



(a) Ground truth (b) Faster RCNN Output (c) RetinaNet Output (d) Yolov3 Output

Figure 4: Comparison between the different output images with the original image.

terms of detection speed. YOLOv3 is 6x faster than Faster RCNN and 2x faster than RetinaNet. Even though Faster RCNN is the slowest, it has the lowest regression loss (0.1772) for boundary box prediction while RetinaNet and YOLOv3 have regression loss 0.3079 and 0.5132 respectively (see Table 1). YOLOv3 being the fastest of all also has the highest mAP of 0.43 while Faster RCNN and RetinaNet have mAPs 0.36 and 0.24 respectively (see Table 1). As seen in (Figure 4) Faster RCNN predicts a minimum number of Boundary Box since we have set a threshold of 0.8. YOLOv3 predictions are coherent with the ground truth or original predictions while RetinaNet predictions are shifted right-skewed.

6 Conclusions

This article presented a comprehensive research on the application of deep learning in the field of medical image detection. We have presented an end-to-end framework for the detection of frame objects in the medical images generated using Endoscopy while leveraging the power of deep neural networks. Each part of our framework is a neural network, which utilizes the real time processing capabilities of modern GPUs. We have experimented with various novel object detection architectures and various image preprocessing techniques on Endoscopy Artifact Detection Dataset. To overcome the problem of small training dataset, we also experimented with various data augementation techniques. In our experiment, we achieved the highest mean average precision (mAP) of 0.43 and mean inference time of 182s for YOLOv3 with a tradeoff in regression loss. Though FasterRCNN achieved the lowest regression loss it is not possible to use it for real-time inference due to the high mean inference time. RetinaNet was behind YOLOv3 and Faster RCNN both in terms of mAP and quality of the predicted boundary box. Thus we conclude that YOLOv3 was a clear choice considering the applicability of our framework in realtime. We demonstrated high-quality performance on real clinical endoscopy videos and images for both intra- and inter-patient variabilities and multimodality. Future work will focus on further improving the object detection network and implementing the entire framework as a single end-to-end trainable neural network.

7 Acknowledgement

We thank the invaluable inputs and suggestions by Prof. Ankit Bansal that were integral for the successful completion of this paper. We also thank the Institute Computer Centre (ICC), IIT Roorkee, for providing GPU workstations required for performing various experiments involved in the paper.

References

[1] Sharib Ali et al. “Endoscopy artifact detection (EAD 2019) challenge dataset.” In: *arXiv preprint arXiv:1905.03209* (2019).

[2] TY Lin et al. “Focal loss for dense object detection. arXiv 2017.” In: *arXiv preprint arXiv:1708.02002* (2002).

[3] Joseph Redmon and Ali Farhadi. “Yolov3: An incremental improvement. arXiv 2018.” In: *arXiv preprint arXiv:1804.02767* (2019).

[4] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks.” In: *Advances in neural information processing systems*. 2015, pp. 91–99.