RetinaNet Application in Bioburden Detection with Physical Layer Simulating Colored Illumination

Jiacheng Lin
Department of Biomedical Engineering
Duke University
Durham, NC 27707
jiacheng.lin@duke.edu

Abstract

Uncleaned tools passing the visual inspection from human auditors at Sterile Processing Department has been a concern to hospitals for years. In this paper, an innovative way of bioburden detection with the help of an automatic imaging system and machine learning algorithms is proposed. Additional to the modified RetinaNet model used in this study, physical layers are also explored, which simulates a color-changeable illumination system. Though the model with physical layer did not perform as well as the one without physical layer, the experiment still pointed out an interesting direction worth further development.

1 Introduction

Sterile Processing Department(SPD) is where all medical equipment is cleaned and sterilized in hospitals. At SPD, the used surgical tools from the OR are firstly manually, hand-washed by technicians. Then they are visually inspected by auditors to make sure all of the bioburden is removed and there is no rust or any contaminants left on the tool. One problem hazed hospitals for years is that some unclean tools may pass the visual inspection by accident and enter the Operating Rooms(OR) in the end. One reason this is happening is that the tools are inspected one by one with the naked eye by the auditors. They usually have to inspect thousands of tools each day, and often in an inconsistent and unfavorable environments, poor lighting for example.

Because of this reason, the visual inspection can be unreliable and may lead to unclean tools being sent to the OR. According to some interviews conducted with doctors and physicians, nearly all surgeons reported to have experience that when the patient was on the bed in OR, they had found the needed tools are not clean and need to hold the surgery waiting for new tools being sent.

To solve this problem, a novel device is proposed to assist technicians to lower the risk of having unclean tools pass the inspection. The device consists of two major subsystems: an automatic imaging system with multiple cameras and a built in computer running convolution neural network(CNN) algorithm to do the bioburden detection based images taken by the imaging system. In this paper, focus is mainly on developing the CNN for the system.

Besides training a CNN on the custom dataset to get an bioburden detector, a simulation of colored illumination is another research interest. The author hypothesis that by collecting images in colored illumination with an intensity sensor, the model could achieve comparable or even better performance than the RGB sensor with white illumination. The hypothesis is made due to the large difference in light absorption between blood spots and silver metallic surfaces.

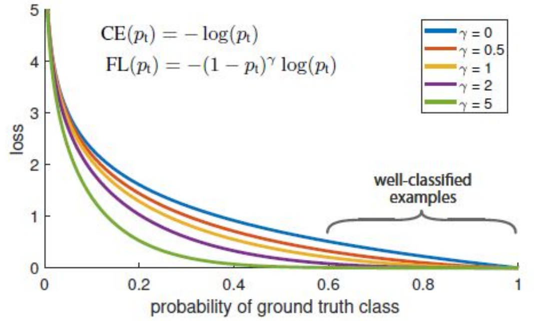


Figure 1: Focal loss can suppress loss of easy negative examples, while keep the loss of hard positive examples

2 Related Work

2.1 Object Detection

Classic object detection like the sliding-window paradigm, in which a classifier is applied on a dense image grid, has a long and rich history. One of the earliest successes is the classic work of LeCun *et al*, who applied convolutional neural networks to handwritten digit recognition [1]. Similar methods were developed and achieved top results on PASCAL for years.

While the sliding-window approach was the leading detection paradigm in classic computer vision, two-stage detectors quickly came to dominate object detection as the deep learning tide started in the early 2000s. The two-stage detectors feature a first stage of generating a sparse set of candidate proposals that should contain all objects while filtering out the majority of negative locations, and a second stage classifies the proposals into foreground classes / background[8]. Along the research in two-stage detectors, the Faster R-CNN introduced Region Proposal Networks (RPN) which integrated proposal generation with the second-stage classifier into a single convolution network framework and hold the crown of object detection for COCO for a decent period of time [6].

Besides two-stage detectors, another category of object detectors merged mainly aiming for speed with a trade-off of accuracy - the one-stage detectors. One-stage detectors do not have any region proposal networks, but instead a usage of anchor boxes which generate over the entire image, at different scales and aspect ratios, to cover all interested features. Early implementations of these detectors suffered from a significant accuracy loss like SSD and YOLO [3, 5]. In 2017, Lin *et al* proposed the focal loss and RetinaNet framework to address the imbalance of foreground and background boxes [2]. They claimed that training with classification loss function of cross entropy loss is inefficient as most locations are easy negatives, which can contribute no useful learning signal. Another negative aspect of classic cross entropy loss is that the easy negatives can overwhelm training and lead to degenerate models. In their work, the focal loss could suppress the loss value of easy negative examples while keep the hard positive examples' loss value relatively the same, as shown in Fig. 1.

The study in this paper chose to work with RetinaNet for its high speed in nature as an one-stage detector, as well as the decent performance it can achieve for dense object detection. Additionally, the decent implements being available in open source community is another consideration for the usage in a medical device.

2.2 Bioburden Detection

Bioburden detection for surgical tools is still a unexplored field at the point this paper is written. Some studies on similar objectives were found and inspired the author regarding collecting images and training for the proposed task [7, 4].

Outside the academic field, the search for commercial solutions and patents shows no similar products been launched or patents been filed. According to an interview with a manager at General Electric, they have been trying to automate the whole process of sterile processing and in which they are

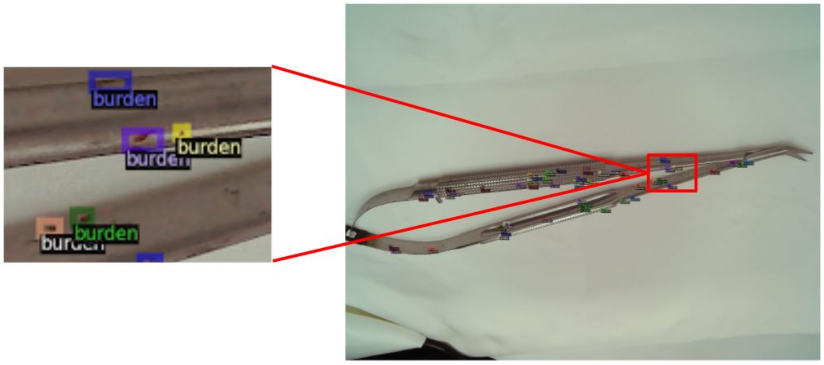


Figure 2: An example of the custom dataset used in this paper

combining some computer vision technologies with their existing robotic systems for surgical tools classification. The proposed device in this paper is innovative and deserve further development to address the concern of unclean tools getting into OR as mentioned earlier.

3 Methods

3.1 Data Collection

The images used in this paper is collected by the author with old surgical tools retrieved from a nearby hospital. To imitate the real uncleaned tools found in SPD, fake blood is applied onto the tools with electrical toothbrush, which can create consistent, small blood spots. After applying fake blood, the tools will be dried for enough time, so the fake blood can congeal to stubborn dark red spots. Then the tools are cleaned by hand with brushes and paper towels to simulate the cleaning process at SPD. The finished tools are placed in a light box for imaging with 3 cameras from above and 45°on right and left sides. The tools are finally thoroughly cleaned and checked before next round of data collection.

In total, 398 images were collected with the process described above and are manually annotated. Across the whole dataset, there are over 12800 blood spots in the annotation. The images are 3264×2448 in resolution and saved as 24-bit JPG files. An example of our custom dataset with annotation can be checked in Fig.2.

3.2 Neural Network Structure

3.2.1 Base RetinaNet

In this paper, the base object detection framework is RetinaNet and used the source code provided by Lin *et al*[2]. Specifically, the configuration used in this paper is the most compact build of RetinaNet, mainly for speed and fast training. The backbone used is ResNet-50, which connects to a 5-level Feature Pyramid Networks(FPN) after 2-5 ResNet block and the P6 in FPN is generated from C5 in backbone via a convolution layer. Then the anchors are generated for all levels in FPN before passing to next stage. The classification and bounding box regression sub-models are built according to origin paper, both are 5-layer convolutional networks. The outputs from two sub-models goes through a Non Maximum Suppression(NMS) layer to generate the final bounding box outputs.

In this paper, all models are trained for 10 epochs for time limitation and fast over fitting on a small dataset.

3.2.2 Anchor optimization

Due to the small-in-size nature of bioburden after initial cleaning step, all objects interested in the custom dataset are significantly smaller than normal objects from COCO dataset. The typical size of bounding boxes in our dataset is 20×20 pixels in 3264×2448 images. When training the model on

Table 1: Anchor configuration after optimization

Anchor		
Parameter	Description	Optimized value after rounding
Size	Initial anchor size on each feature level	32, 64, 128, 256, 512
Strides	How network strides over features	4, 8, 16, 32, 64
Ratios	Aspect ratios of anchors	0.7, 1, 1.43
Scales	Scaling factors for each anchor location	0.2, 0.3, 0.4, 0.5

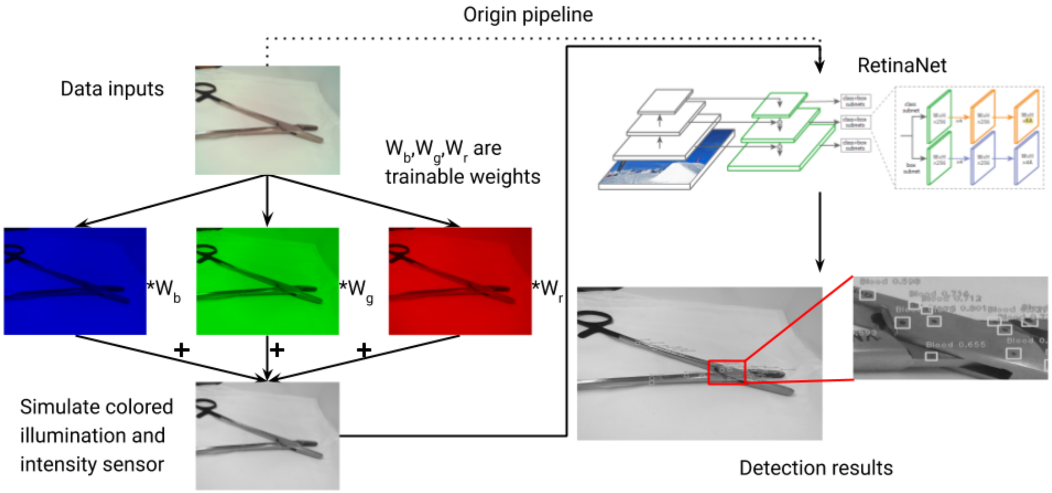

Origin pipeline
Data inputs
RetinaNet
W_b, W_g, W_r are trainable weights
$*W_b$
$*W_g$
$*W_r$
+
+
+
Simulate colored illumination and intensity sensor
Detection results

Figure 3: An example of the custom dataset used in this paper

our dataset with default anchor configurations, the model failed to recognize any objects because of too large anchor boxes.

The anchor configuration used in this paper is obtained from optimization methods introduced by Zlocha *et al* in 2019[9]. The final optimization results are listed in Table 1.

3.2.3 Physical Layer

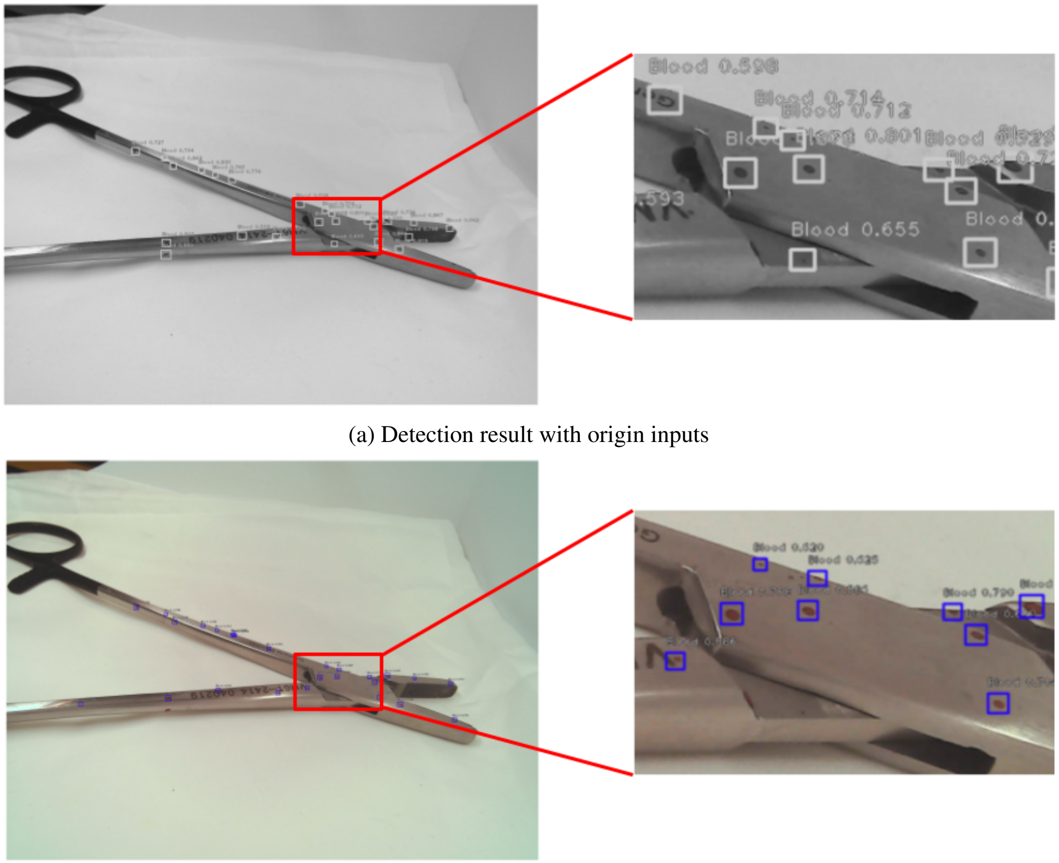
The final physical layer chosen for this paper is a weighted sum layer, simulating the colored illumination during data collection. The overall structure of whole pipeline is shown in Fig 3.

As Fig 3 shows, the physical layer operates by multiplying each channel of the input image with a trainable weight and then sum all channels together. This yields a single channel intensity image. The idea behind the design is to simulate collecting data with colored illumination and a intensity sensor to optimize detection while improve speed by reducing input data size. The weights are trainable and is optimized with the CNN weights during training.

4 Results

The evaluation results on separated test set of 25 images are listed in Table 2. Detection examples for model without physical layer(Fig 4a) and with uniform weighted sum(Fig 4b) are provided in Fig 4.

As can be seen in Table 2, the model with physical layer of weighted sum simulating colored illumination achieved an AP of 0.513, while model without physical layer achieved 0.63 for AP. And the model without physical layer takes around $\sim$ 140ms to do inference on one image than the model with physical layers. This proves how input size influence the detection speed. The physical layer introduces a trade off between inference speed and accuracy that needs further experiments and development.



(a) Detection result with origin inputs

(b) Detection result with physical layer

Figure 4: Detection examples after 10 epochs training

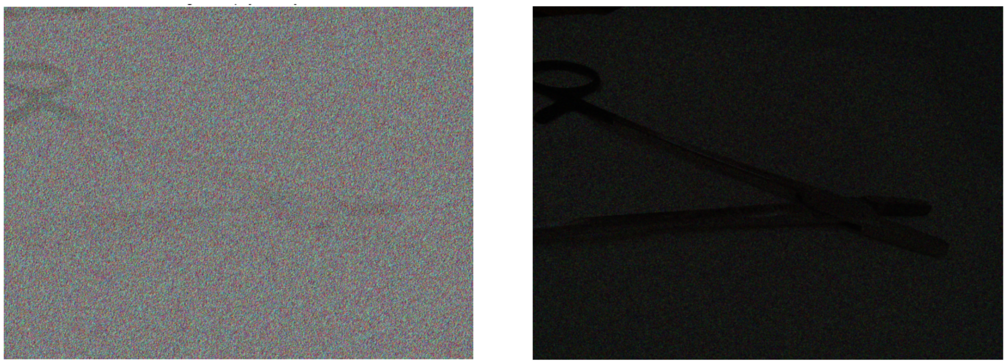
Table 2: Training results on test set with 10 epochs

Model	Average Precision	Inference time
RetinaNet(ResNet50 + FPN)	0.630	$\sim$ 350ms
RetinaNet + Weighted sum (uniform weighted)	0.513	$\sim$ 210ms
RetinaNet + Weighted sum (pixel-wise weighted)	0.165	$\sim$ 210ms
RetinaNet + pixel-wise scaling	0.167	$\sim$ 360ms

And after training, the changes in weights are different across three color channels. Decrease in red channel weight and increase in both blue and green channel were observed. These changes matches the hypothesis that red blood spots will have higher contrast when illuminated with colored illumination. And our team will test different illumination color with physical hardware to validate the findings in this paper.

5 Discussion

One major goal for this paper is to investigate if physical layers can improve model performance. For the decrease in average precision, one explanation could be the result of summing channels together, which suffers from some information loss about the interested objects. Since the image is turned into a monochrome gray scale image, the information of color is lost and intensity information is kept. The model might fail to achieve such high accuracy with only intensity data. And the contrast increase from introducing physical layer could be minor for blood spots detection.



(a) Example image after model 3 trained physical layer(b) Example image after model 4 trained physical layer

Figure 5: Trained weights applied to example image for model 3 and 4. Weights are randomly initialized.

When looking into the zoomed vision in Fig 4, we can see in Fig 4a the model failed to detect some very small spots. And in Fig 4b, we can see those small spots are detected with good confidence score. But in the same time, the label letter *G* at top left corner is also labelled as a blood spots. One possible reason is that the model with physical layer detect blood spots based on the contrast for a certain area. In this way, the model tends to label patterns have distinct contrast difference with surroundings and thus labeled many non-blood objects as blood. In fact, if the device would be super sensitive and aims for low false negative rate, the colored light might be better to make sure no spot is missed.

And regarding the model 3 and model 4 in Table 2, those are other physical filters author tried and failed. The model 3 features a pixel wise weighted sum, which has $width \times height \times channels$ trainable weights. And model 4 features a pixel wise scaling. The only difference between model 3 and 4 is that in 4, channels are not summed together and that is the reason model 4 took longer time to inference. These models are made with an assumption that certain sensor array can achieve best sensing for the model. The trained weights can be used as a reference for sensor design in further development. Figure 5 shows trained physical layers applied to example images. As can be seen in the figures and the accuracy in Table 2, the training was not successful. Also, very minor changes were observed on the weights in model 3 and 4. Our thought is that 10 epoch training is not effective enough for this many weights. Or, the loss functions are not designed to optimize these weights and the learning rate might decay too fast to really improve these weights.

Acknowledgments

This paper is written as part of the requirement of final project for *BME548L Machine Learning and Imaging* given by Dr. Roarke Horstmeyer at Duke University. His guidance is appreciated. Same appreciation goes to Colin Cooke, who provided generous help during the model debugging and physical layer design.

And finally, thanks the SPDream team members: Moriah Garcia, Nicole Belisle, Neil Gupta and Amanda Huang for proposing such a wonderful idea with the author and annotating numerous images.

Please notice: This study is being developed and intended for patenting. Please do not share without permission of the author.

References

[1] Yann LeCun et al. “Backpropagation applied to handwritten zip code recognition”. In: *Neural computation* 1.4 (1989), pp. 541–551.

[2] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[3] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[4] Balakrishnan Ramalingam et al. “Vision-Based Dirt Detection and Adaptive Tiling Scheme for Selective Area Coverage”. In: *Journal of Sensors* 2018 (2018).

[5] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[6] 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.

[7] Xian Tao et al. “Automatic metallic surface defect detection and recognition with convolutional neural networks”. In: *Applied Sciences* 8.9 (2018), p. 1575.

[8] Jasper RR Uijlings et al. “Selective search for object recognition”. In: *International journal of computer vision* 104.2 (2013), pp. 154–171.

[9] Martin Zlocha, Qi Dou, and Ben Glocker. “Improving RetinaNet for CT Lesion Detection with Dense Masks from Weak RECIST Labels”. In: *arXiv preprint arXiv:1906.02283* (2019).