卒 業 論 文 概 要 書

Summary of Bachelor's Thesis

Date of submission: 07/24/2023 (MM/DD/YYYY)

学科名 Department	Major in Computer Science and Communications Engineering	氏名 Name	Kein Yamada	指 導 教 員 Advisor	渡辺 裕 ㊞
研究指導名 Research guidance	Research on Audiovisual Information Processing	学籍番号 Student ID number	1W19CF09-1		
研究題目 Title	Application of Contrast Reduction for Accuracy Consistency in Object Detection				

1. Introduction

In recent years, the utilization of artificial intelligence (AI) in video and image processing has been growing with continuous demands. Following the improvement of AI for image recognition, computer vision is becoming more capable of image and video processing. Thus, images and videos are starting to be more applicative for these purposes. If the information of images can be reduced to satisfy the purposes above, an efficient transmission can be achieved. Furthermore, it has been revealed that images required for computer vision can be smaller, compared to ones for human [1]. By reducing the information of the image, we attempt to restrain the entropy.

Entropy in image processing measures the level of disorder within an image, which is determined by the frequency of each pixel value. In this research, the entropy is suppressed by placing a limit on pixel values that could potentially appear. By narrowing down the pixel values to a certain range, the contrast of the image is reduced. Hence, we decided to adopt the alternation of pixel value range as an entropy reduction method, which will be denoted as contrast reduction in this paper.

2. Related works

YOLOv5 [2] is a modified object detection model, belonging to the You Only Look Once (YOLO) family [3]. YOLOv5 is built to have a quick process of training and is highly suitable for users that insist on installing computer vision technologies to their proposed mechanism. The model consists of 4 main versions named small (s), medium (m), large (l), and extra-large (x). In terms of model structure, YOLOv5 consists of 3 components called Backbone, Neck, and Head, making it a single-stage object detector. As a feature of single-stage object detector, the neural network is passed through once and predicts the object's coordinates, size, and classification simultaneously [4].

3. Proposed method

In this section, we propose a method of lowering the entropy by reducing the contrast of the image. In this research, images are treated as 3 channel images with a bit depth of 8. Since the images have a bit depth of 8, pixel values range from 0 to 255. Depending on an assigned pixel value, a pixel depicts a certain intensity of the color shade. The process of contrast reduction requires the modification in pixel values. The values are changed by the limitation of the pixel value range, conducted by the following equation,

$$O(w,h) = \frac{\alpha}{256} I(w,h) + \frac{256-\alpha}{256WH} \sum_{w=1}^{W} \sum_{h=1}^{H} I(w,h). \quad (1)$$

I(w, h) and *O(w, h)* represent the pixel values of the input image and output image. Symbol α refers to the target range of the pixel value. For example, in the case of limiting the pixel value range to 128, α should be equal to 128 as well. W and H stand for the

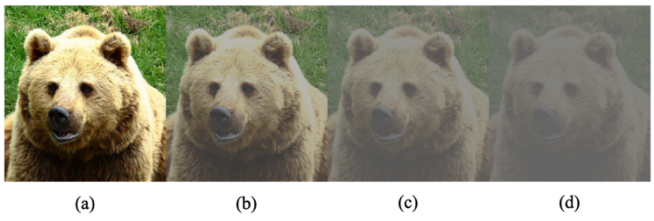


Fig. 1. Visualization of images in different pixel value range, (a) range of 256; (b) range of 128; (c) range of 64; (d) range of 32.

Fig. 2. Overview of contrast reduction and training process.

width and height of the image. The expression after the + sign indicates the average pixel value of the entire input image. Moreover, this expression aids in adjusting the limit range to optimally use the pixel values close to the original image. With (1), an image can be successfully converted to have a smaller range of pixel values.

The range downscaling directly correlates to the reduction of entropy within the images. This is because the randomness of the image's colors per pixel gets smaller with range limitation. For instance, if the range limit is shrunk to express 128 pixel values, usable colors for image representation is halved. With fewer colors used for expression, the color intensity or the shade within the images are less dynamic. Therefore, the texture of an image will be more simplified, resulting in an entropy decrease.

4. Experiment

To validate our proposed method, we examined the relationship between the accuracy of object detection and the entropy of the images. The datasets used for training and validation are COCO 2017 Train Images and COCO2017 Val Images [4]. As mentioned in the proposed method, images of the datasets are converted by contrast reduction to pixel value ranges of 256, 128, 64, and 32, as shown in Fig. 1. In Fig. 2, the overview of the experiment is demonstrated. For the object detection model, YOLOv5m is used for training and validation. Additionally, images of the converted datasets are utilized for the process of object detection. For each pixel value range, YOLOv5m has been trained on 100 epochs. After completing the train on YOLOv5m, the output results are examined, and the best values are used for comparison with different ranges of pixel values. For the evaluation, mean Average Precision (mAP) values are depicted and split into mAP50 and mAP50-95. The entropy of an image is calculated by the concept of Shannon entropy. The equation for entropy is represented as

$$H = -\sum_{i=0}^{255} p_i \log_2 p_i. \qquad (2)$$

Using (2), the entropy is calculated for each image in the validation dataset. Since the COCO2017 Val dataset includes 5000 images, the mean of 5000 entropy values is used for the result. For the object detection model, YOLOv5m is used for training and validation.

TABLE I
DETECTION ACCURACY AND ENTROPY FOR EACH PROCESS

Process	Entropy	mAP50-95	mAP50	mAP75
PVR256	7.216	0.445	0.637	0.484
PVR128	6.232	0.447	0.640	0.485
PVR64	5.251	0.447	0.639	0.484
PVR32	4.285	0.443	0.635	0.481
PVR16	3.346	0.438	0.630	0.474
PVR8	2.448	0.423	0.614	0.456
PVR4	1.594	0.385	0.573	0.411
Sobel	6.103	0.369	0.543	0.399
Canny	0.617	0.312	0.469	0.335
LoG	3.985	0.337	0.503	0.361

Shown above in Table 1, results of the mAP values and entropy are presented with the corresponding processes. PVR is short for pixel value range, meaning process PVR64 stands for a pixel value range of 64. LoG is an abbreviation for Laplacian of Gaussian. Results indicated that values for mAP50-95 stayed consistent throughout the fall in pixel value range until it reached 32. Furthermore, values for mAP50 and mAP75 held great stability in accuracy, barely changing in numbers. Similarly, values in entropy also decreased as the pixel value range became smaller. Our experiment successfully conveyed that the pixel value range of an image can be downscaled to 32 while providing an object detection performance close to the original. Hence, the output results of Table 1 revealed that the detection accuracy can be maintained while dropping the entropy with contrast reduction.

5. Conclusion

The outcomes of this research have solidified the use of contrast reduction as a reliable entropy reduction method for omitting excess textures within the image without greatly decreasing the performance of object detection. Consequently, the entropy of an image can be decreased to be applied to computer vison-based image recognition. For future research, we plan on finding a better contrast reduction method and consider a way to connect this research to image coding.

6. Reference

[1] H. Choi, and I. V. Bajić. "Scalable Image Coding for Humans and Machines," *IEEE Transactions on Image Processing*, vol. 31, pp. 2739-2754, Mar. 2022, DOI: 10.1109/TIP.2022.3160602.

[2] G. Jocher, A. Chaurasia, A. Stoken, *et al.*, *ultralytics/yolov5: v7.0 – YOLOv5 SOTA Realtime Instance Segmentation*, version v7.0, Nov. 2022, DOI: 10.5281/zenodo.7347926.

[3] J. Redmon, S. Divvala, R. Girshik, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788, 2016. DOI: 10.1109/CVPR.2016.91.

[4] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft COCO: Common Objects in Context," *Computer Vision – ECCV 2014*, vol. 8693, pp.740-755, Sep. 2014. DOI: 10.1007/978-3-319-10602-1_48.

2023 Bachelor Thesis

Application of Contrast Reduction for Accuracy Consistency in Object Detection

School of Fundamental Science and Engineering,

Major in Computer Science and Communication Engineering

Submission Date: July 24th, 2023

Kein Yamada
(1W19CF09-1)
Supervisor: Prof. Hiroshi Watanabe
CSCE, Waseda University

Contents

Chapter 1

Introduction

1.1 Research Background

Artificial intelligence (AI) has been showing growth and popularity many tasks of video and image processing. Following the improvement of AI for image recognition, computer vision is also becoming more capable of image and video processing. As an example, the combination of AI with object detection [1] has been trending in recent years. There are many real-life applications of object detection such as surveillance system [2], face recognition [3], autonomous driving [4], medical anomaly detection [5], animal classification [6], and more.

Currently, most of these detection tasks utilize clear video and images with high quality information content. If the information within videos and images can be reduced to satisfy the purposes above, an efficient transmission can be achieved, contributing to a quicker processing time. Furthermore, it has been explained that information content in videos required for computer image recognition can be smaller than ones for human [7]. Since the videos for machines do not require excessive information for detecting objects, theoretically the AI should behave similarly for the case of images. In order to reduce the information content in media files such as images, an image degradation method can be considered.

1.2 Related Works, Problems, and Research Purpose

For an AI to precisely and accurately detect the objects, the machine needs a training model to learn from, which is called object detection model. In recent years, there are many variants and modified versions of the object detection model, each having its unique application purposes. The 4 major ones are known to be Region-based Convolutional Neural Network (R-CNN)

[8], Fast Region-based Convolutional Network method (Fast R-CNN) [9], Single-Shot Detector (SSD) [10], and You Only Look Once (YOLO) [11]. In this thesis, YOLO will be the main model discussed since it is utilized and explained in the experiment.

Entropy in image and video processing have a different definition as the one used for thermodynamics. Similar to the one defined in statistical mechanics, the entropy measures the randomness in images or frames for videos by referring to the frequency of each pixel values. Generally speaking, high entropy resembles a complex texture within the image with frequent change in color usage. Considering a way to decrease the value of entropy, one can attempt the use of edge detection filters. For instance, Sobel operator [12], Canny edge detection [13], and Laplacian of Gaussian (LoG) filter [14] are major edge detection filters that can successfully decrease image entropy. However, these filters also provide a great decrease in accuracy, discussed later in Chapter 4 Experiments.

With a detrimental issue of detection accuracy dropping when images are processed by edge detection filters, a different approach is needed for lowering the entropy but also maintaining the accuracy. As a method that fulfills these two requirements, this thesis proposes the suppression of entropy through the limitation of pixel value frequencies within an image. Placing a restriction on the range of pixel values directly reduces the contrast of an image, which should also decrease the entropy. Surprisingly, there are roughly any researches that depict contrast reduction as a way of entropy reduction. Hence, pixel value range management will be used as a proposal for restraining the entropy. Such process will be referenced as contrast reduction in this thesis.

1.3 Thesis Outline

The outline of the thesis is followed as below:

Chapter 1: The overview of object detection and its application is provided. The chapter also explains the significance of reducing the information within media files, convey an issue about entropy reduction through edge detection filters, and propose a new method of lowering entropy by contrast reduction.

Chapter 2: Research and object detection models related to this thesis are presented. This chapter will describe the structure and purpose of the first YOLO model. Afterwards, features of YOLOv3 and YOLOv5 are explained to point out the modifications made for detection improvement.

Chapter 3: The proposed method of this thesis is clarified. In this chapter,

the general understanding of images and pixel values is introduced. Later, the process of contrast reduction and the equation used will be expressed.

Chapter 4: Experimental procedure, results, and discussion are demonstrated. Experiment settings and calculation of obtained results are also mentioned.

Chapter 5: Thesis is concluded with an achievement summarized from Chapter 4 about contrast reduction and possible changes and adjustments for the future are discussed.

Chapter 2

Related Works

2.1 You Only Look Once (YOLO)

You Only Look Once (YOLO) is one of the object detection model widely used nowadays, currently provided in many different versions. YOLO is a model first showcased in a research paper called *You Only Look Once: Unified, Real-Time Object Detection*, published in 2015. When YOLO was introduced, the model showed a great impact and uniqueness as it served as the first object detection architecture to predict the size, location, bounding box, and the classification of the object at once. A model with this feature is now known as a phrase called one-stage object detection model. On the other hand, R-CNN and Fast R-CNN, or other models that requires multiple steps to execute all the object detection process are called two-stage object detection model. Unlike the two-stage object detection models, the strength of YOLO lies in the functionality of attempting the object detection predictions all at once, as shown in Figure 1. Due to this feature, YOLO has a rapid processing time, making the model suitable for real-time detection.

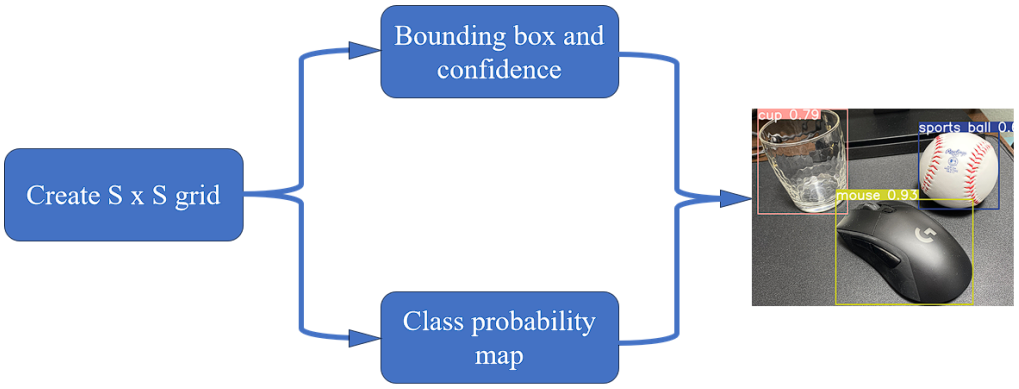


Figure 1. Overview of the YOLO mechanism.

The first version of YOLO consists of 4 big steps for completing the task of object detection. In the first step, YOLO uses residual blocks to divide an image into a $S \times S$ grid with cells of equal shape. Whenever the objects are present in the cells or are covering them, the corresponding grid cells attempt detection, location prediction, and classification prediction. The second step involves the bounding box regression, which is a technique utilized for making a prediction about the object's location through model training. This step enables the prediction of bounding box surrounding the object, where the attributes are given as height, width, center, and object class. For the third step, a concept called Intersection over Union (IoU) is applied. The equation of the IoU is given as

$$IOU = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{1}$$

where the Area of Overlap is the intersection of ground truth area and predicted area and Area of Union is the union of ground truth area and predicted box area. IoU provides a general understanding of how the bounding boxes and grid boxes overlap, which is helpful with achieving a bounding box fitting for the object and removing the rest that are unnecessary. The removal of redundant bounding boxes is achieved by the IoU threshold, which can be manipulated by the user as an input for limiting the predictions made by grid cells with low IoU. Finally, the fourth step exploits a computer vision method called Non-Max Suppression (NMS). Unfortunately, there are occasions where the IoU threshold will not be enough to diminish the insignificant bounding boxes. NMS helps the extraction of the noises created by overlapping bounding boxes, leaving only the boxes with the best detection score.

2.2 YOLOv3

Ever since the release of the first version of YOLO, many versions are created and published. YOLOv3 [15] is the last object detection model from the YOLO family that has been published as a research in 2018 by the official creators, Redmon et al. This model is based off YOLO9000 [16], known to be the second version of YOLO created by the same publishers around the end of 2016. There has been major improvements from YOLO to YOLO9000 through the addition of many significant features. The new functions are batch normalization, higher resolution classifier, anchor boxes, fine-grained features, multi-scale training, and the Darknet-19 architecture. These advancements structuralized YOLO9000 to be better, faster, and stronger as mentioned in their research title. YOLOv3 continued on with the YOLO9000

features but also made some changes for incremental improvement. While it is true that YOLO9000 is faster and more efficient than YOLO, it is not enough to compete with the best object detection algorithms that insist on greater accuracy and quick detection. The modifications made in YOLOv3 are bounding box predictions, class predictions, Feature Pyramid Networks (FPN), and the Darknet-53 network. For bounding box predictions, logistic regression is now used for predicting the objectness score, a score that quantifies the degree at which the object detector successfully guesses the correct class and location of objects. Class prediction involves the utilization of logistic classifiers in YOLOv3 instead of softmax, enabling multi-label classification. YOLOv3 adopted a system similar to the FPN for feature extraction from 3 different scales of bounding box prediction. This helps with achieving more information needed for improved detection, benefited from the scales. Finally, Redmon et al. moved on to a new network for YOLOv3, from Darknet-19 to Darknet-53. The convolutional layers has been increased from 19 to 53 where several layers added serve the purpose of shortcut connections. Refinements made in Darknet-53 provides an efficient use of the GPU, supporting the overall detection. Thus, YOLO9000 has been optimized to YOLOv3, making the model great use for general object detection tasks.

2.3 YOLOv5

The official production of the YOLO model has been paused at YOLOv3. However, there has been many upgrades made to YOLO's object detection ever since the release of the last official model. One of those improved models is YOLOv5 [17], object detection model created by Ultralytics in 2020. YOLOv5 is designed to have a rapid training processing with many customizable options for the users. Ultralytics provides the model in 5 different main versions that are called nano (YOLOv5n), small (YOLOv5s), medium (YOLOv5m), large (YOLOv5l), and extra-large (YOLOv5x), as shown in Figure 2. The 5 pretrained models follow a specific trend of object detection evaluation and processing time. If one insists on acquiring the result quickly, regardless of its accuracy and precision, YOLOv5n should be used since it has the smallest parameter among the 5 models, giving the shortest detection process. On the other hand, if the emphasis is placed more on the detection evaluation than the processing time, YOLOv5x with the largest parameter should be considered.

Furthermore, YOLOv5 consists of many types of data augmentation techniques. The augmentation methods include mosaic augmentation, copy-paste

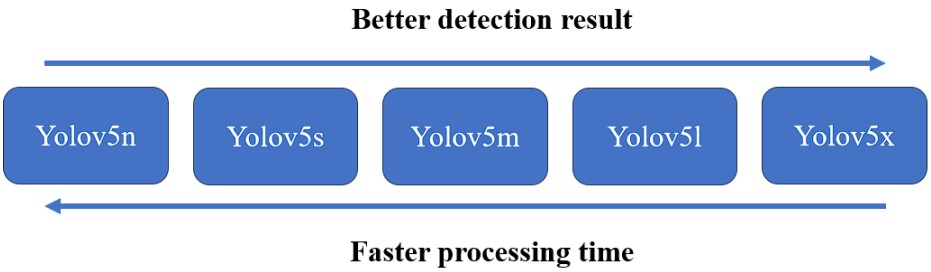


Figure 2. Diagram showing the strength and weakness of Ultralytic's pre-trained YOLOv5 models.

augmentation, random affine transformations, MixUp augmentation, albumentations, HSV augmentation, and random horizontal flip. Mosaic augmentation merges 4 training images into 1, helping out the object detection models with image size scaling and better processing. Other augmentation methods create more sample data images for the object detection model to train on. Each techniques also have a specific use, making it easy for the users to achieve a data augmentation that satisfies their desired dataset and training objective. Due to these data augmentation methods, YOLOv5's generalization is greatly improved and the risk of quick overfitting is reduced.

YOLOv5 has a complex model structure but the architecture can be separated into three main parts just like the other single stage object detectors. These parts are called backbone, neck, and head. Backbone acts as a pre-trained network that helps extract feature of multiple resolutions. The neck attempts feature pyramid extraction, supporting the model's generalization. The neck is also the connection necessary for linking the backbone and the head. Model head is in charge of producing the final output through last stage operations. It is what adds anchor boxes onto feature maps to write out the classes, objectness scores, and bounding boxes. With this model structure, YOLOv5 make dense predictions to achieve better results for object detection. As shown below in Figure 3, CSP-Darknet53 is used for backbone, PANet for neck, and the YOLO layer from the YOLOv3 for head.

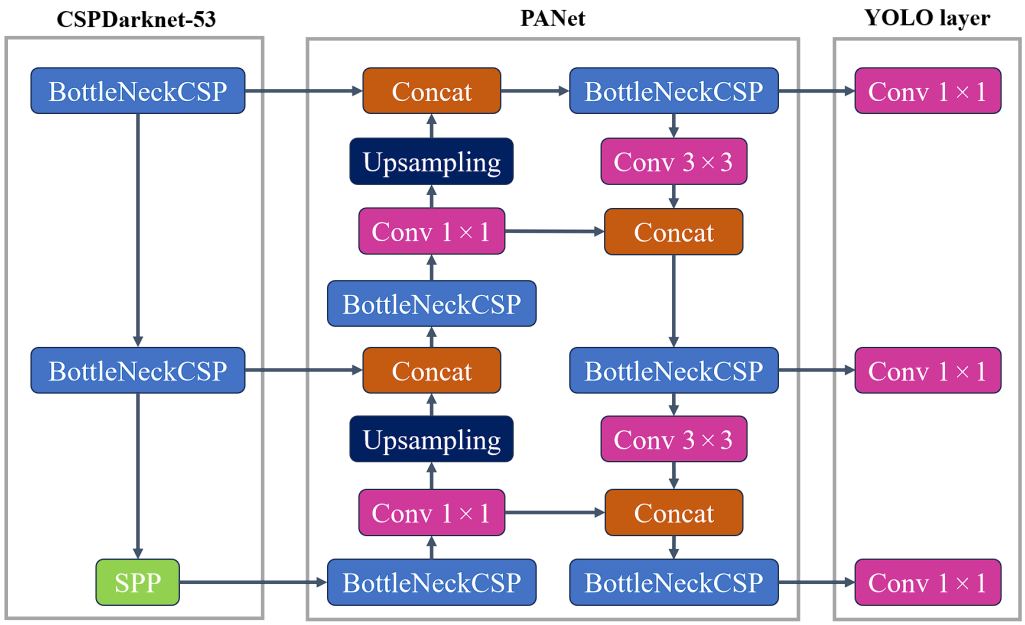


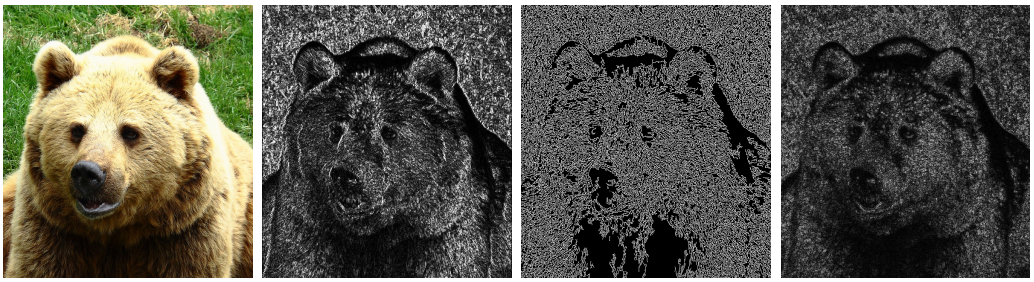
Figure 3. Organized diagram representing the YOLOv5 model structure.

Chapter 3

Proposed Method

3.1 Edge Detection Filters

Originally, this whole research started with applying edge detection filters to the images used for object detection. Before switching to contrast reduction, edge detection filters were the main method for entropy reduction. The edge detection filters used are Sobel operator, Canny edge detection, and the LoG filter. These edge detection filters are operated onto images from the Microsoft COCO dataset [18], developed by Lin et al. for the purpose of object recognition. The images filtered by edge detection can be observed in Figure 4. After processing the images, the filtered images are inputted into YOLOv3 as train and validation dataset.



(a) Original (b) Sobel filter (c) Canny filter (d) LoG filter

Figure 4. Alignment of the original image and the edge filtered images. The original image is from the COCO2017 dataset [18].

Unfortunately, there are no checkpoints for YOLOv3 about the COCO 2017 dataset, making it difficult to compare the achieved results of the models trained on edge detection filter processed images. Therefore, the experiment will be changing the object detection model to YOLOv5, since Ultralytics provides the pre-trained checkpoints of the YOLOv5 models on their document page.

3.2 Contrast in Digital Images

3.2.1 Overview of Digital Images

Digital images are referred to as electronic photos that are either scanned from physical documents or captured from a certain scene. Every digital image is composed of smallest quantitative portions in the image called picture elements (pixel). Pixels are also capable of representing a specific color or gray level intensity, meaning that they are basically tiny colored dots. These pixels are more distinguishable when the image is printed out in low resolution or zoomed in, as it can be seen in Figure 5.

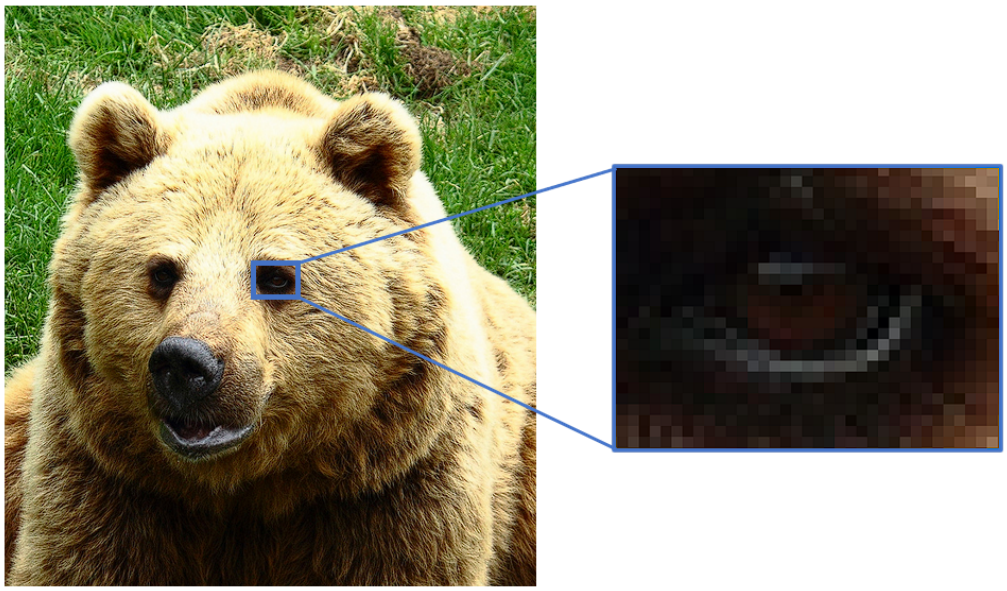


Figure 5. A picture of the bear from COCO2017 dataset [18] zoomed in to show the image pixels.

Generally speaking, digital images is categorized into two major types: color and black and white. Looking at black and white images or grayscale images, these images are presented by different shades of gray. Usually, the gray levels have a range of 256 different grays, as shown in Figure 6. In this case, the representation of 256 levels of shades in black and white pixels can be stored in an 8 bits memory. For color images, the pixels have 3 numbers that resemble the color intensity levels of red, green, and blue, sometimes referred to as RGB channels. If we assume that the color images are displayed in primary colors each having 256 levels, the color pixels can be kept in 24 bits of memory.

Figure 6. Visualization of a grayscale spectrum that represents the black and white images. The scale starts with 0 and ends with 255, which are the shade values expressed in 8-bit integer format[19].

3.2.2 Contrast Reduction Method

Seeking for a different method to lower the entropy, contrast reduction will be introduced in this subsection. From this point onwards, digital images will be treated in a form of 3 channel images where pixel values are formulated by a bit depth of 8. Considering a bit depth of 8, pixel values will have a range of 0 to 255 for all 3 channels. In the task of contrast reduction, the pixel values of the digital image will be altered. These values are modified by setting a limit on the pixel value range, separated into two steps, pixel value limitation (PV limitation) and adjustment. The process of contrast reduction can be observed in Figure 7.

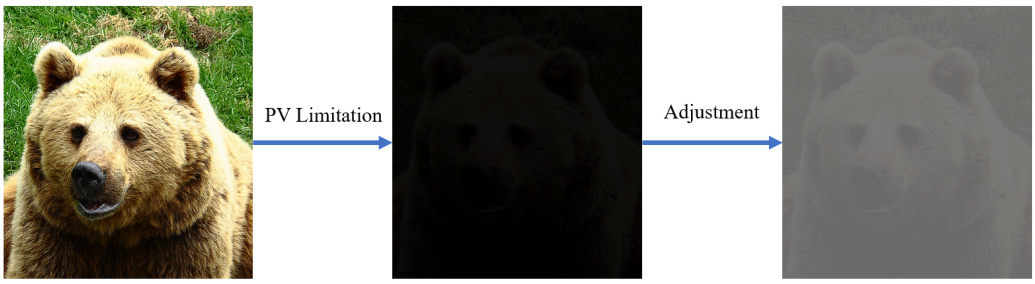


Figure 7. Visualization of the contrast reduction process. The original bear image is from the COCO2017 dataset [17].

As shown in Figure 7, the original image is used as the input in which the input values are reduced by a scale factor of 1/16. In this case, the digital image produced from PV limitation is too dark to see, meaning the adjustment is necessary to make the image brighter and visible. After applying the adjustment process step, the image is more clear because the pixel range is

replaced to fit the colors frequently used in the original image. The whole process of contrast reduction can be described by equation 2.

$$O(w,h) = \left\lfloor \frac{\alpha}{256} I(w,h) \right\rfloor + \left\lfloor \frac{256-\alpha}{256\,WH} \sum_{w=1}^{W} \sum_{h=1}^{H} I(w,h) \right\rfloor \tag{2}$$

where:

$I(w,h) =$ original image, represents the pixel values of the input image
$O(w,h) =$ processed image, represents the pixel values of the output image
$\alpha =$ target range of pixel values, plug in 32 for a range of 32
$w =$ x-coordinate of the digital image
$h =$ y-coordinate of the digital image
$W =$ width of the digital image
$H =$ height of the digital image.

The portion $\left\lfloor \frac{\alpha}{256} I(w,h) \right\rfloor$ corresponds to the PV limitation step mentioned previously. The other portion $\left\lfloor \frac{256-\alpha}{256\,WH} \sum_{w=1}^{W} \sum_{h=1}^{H} I(w,h) \right\rfloor$ corresponds to the adjustment step also mentioned earlier. Truncation is operated on both portions because the images will not be represented in the specified pixel value range if the pixel values are rounded up or left in decimals. While this equation looks complete, it is actually a simplified version as the equation is for a single color channel, meaning this will work for grayscale images. For color images that use an RGB channel, equation 2 needs to be conducted for each channel. This process can be expressed as:

$$O(w,h) = \left\lfloor \frac{\alpha}{256} I(c,w,h) \right\rfloor + \left\lfloor \frac{256-\alpha}{256\,CWH} \sum_{c=1}^{C} \sum_{w=1}^{W} \sum_{h=1}^{H} I(c,w,h) \right\rfloor \tag{3}$$

where c represents the selected channel (R, G, or B for RGB images) and C is the total number of channels, which will be 3 for this research. Thanks to equation 3, the input image goes through contrast reduction to generate the output image with a pixel value range narrowed down.

When a limitation is placed on the pixel value range, it also contributes to reducing the entropy within the images. As mentioned earlier in the Introduction, entropy is a measurement of the randomness of the image's colors per pixel. If the range gets smaller, then the randomness of different

pixel values appearing will also get suppressed. Consider an example where the pixel value range is set from 256 to 128. In this case, the colors that can be used for representing the digital image is halved. Having less colors to work with, color intensity or the shade levels within the images become less dynamic. Thus, the output image will have a simpler texture compared to its original image, leading to a decrease in entropy.

Chapter 4

Experiments

4.1 Experiment Details

With the aim of confirming that the proposed method works properly, the correlation between the accuracy of object detection and the entropy within images will be reviewed. For experimental flow, contrast reduction method explained earlier is applied to the images first. Then, the processed images are used for training and validating for YOLOv5, as shown in Figure 8.

Figure 8. The general flow of the experiment is demonstrated. The original bear image is from the COCO2017 dataset [18].

As it is already revealed in Chapter 3 Proposed Method and Figures 4, 5, and 7, this experiment utilizes the COCO dataset. For the training process, COCO 2017 Train Images are used. For the validation process, COCO 2017 Val Images are applied, which outputs the results examined later in Experimental Results and Discussion. In the next step, all of the images in the COCO 2017 dataset will be converted by the contrast reduction method. After the conversion, there will be 7 datasets with different pixel values ranges, which are 256, 128, 64, 32, 16, 8, and 4. The sample images with various pixel value ranges are shown in Figure 9. These images will be used for training object detection model YOLOv5m and validating the detection performance. Also, the model is trained on 100 epochs per pixel value range.

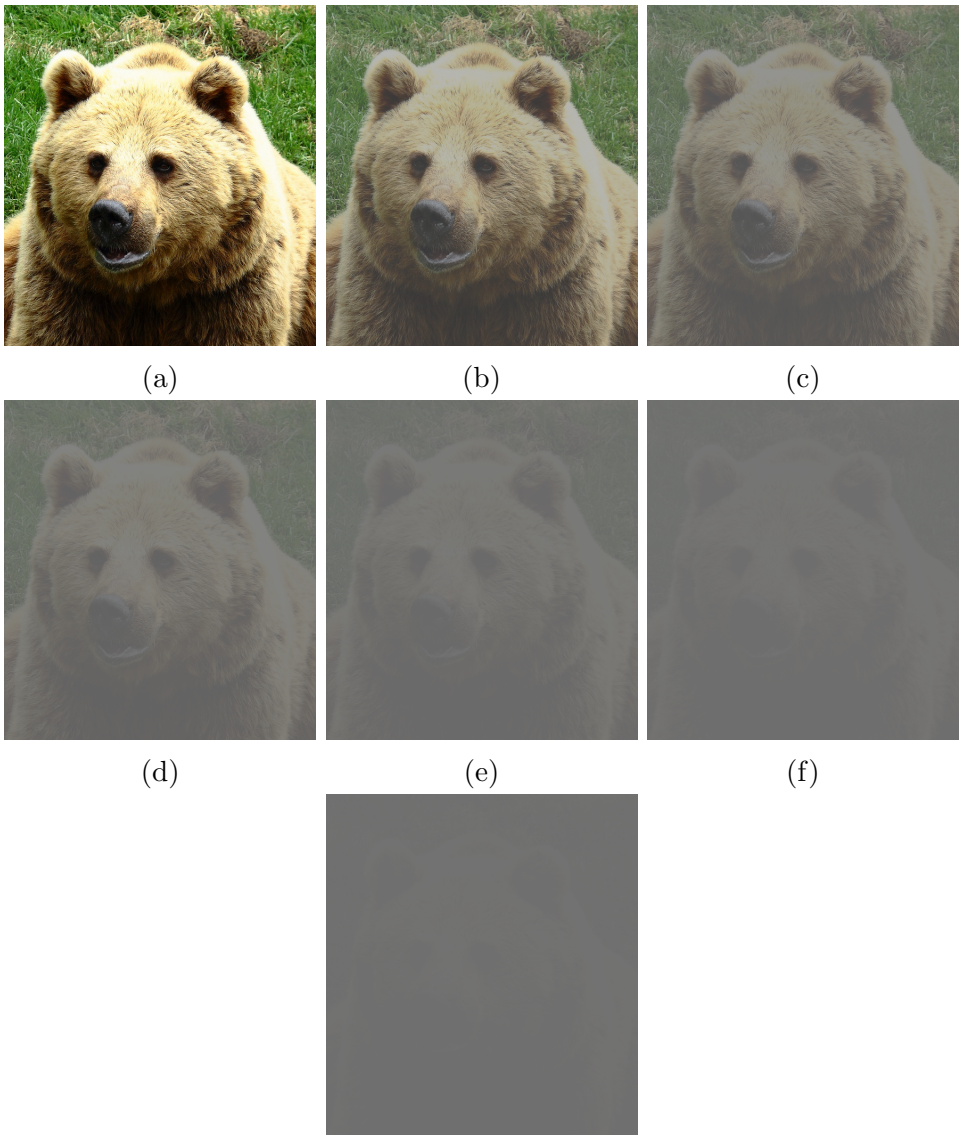


Figure 9. Visualization of images in different pixel value range, (a) range of 256; (b) range of 128; (c) range of 64; (d) range of 32; (e) range of 16; (f) range of 8; (g) range of 4. The original image is from the COCO2017 dataset [18].

4.2 Evaluation Method

In order to evaluate the proposed method, mean Average Precision (mAP) and entropy values of each image processes will be compared. Therefore, 13 processes will be examined, as it can be seen in Table I. For mAP evaluation, values categorized as mAP50, mAP75, and mAP50-95 will be depicted. The number after mAP refers to the IoU threshold of the Average Precision (AP) values used to calculate the mAP. Within these three mAP labels, mAP50-95 is the most important evaluation factor for object detection. This is because mAP50-95 is an average mAP value of the mAP values observed within the IoU threshold range of 50% to 95%, which provides a less biased accuracy to account on. It is also frequently used as an evaluation benchmark for different object detection models. To find the entropy of the images, the concept of Shannon entropy will be used [20]. From Shannon entropy, the equation for entropy is given as:

$$H = \sum_{i=0}^{255} p_i \log_2 p_i \tag{4}$$

where:

H = entropy of an image
i = specific pixel value from 0 to 255
p_i = probability associated with pixel value i.

By utilizing equation 4, the average entropy is calculated for evaluation reference. The average is gained from taking the mean of the entropy sum of 5000 images found in COCO2017 Val dataset.

4.3 Experiment Results and Discussion

Table I reveals the detection accuracy and entropy of the contrast reduction method for 7 different pixel value ranges and the 3 edge detection filter methods. To fit the table size, pixel value range is abbreviated as PVR, meaning PVR64 found in Table I refers to the contrast reduction changing the pixel value range from 256 to 64. Observing the results achieved in Table I, there are many significant matters that can be discussed. First, looking across PVR256 to PVR4, the entropy has successfully dropped from 7.216 to 1.594, which is a 77.9% decrease. While the entropy cannot be reduced to the point of Canny filtered images, even PVR4 with the lowest accuracy

has a greater mAP values than the 3 edge detection filters. The results also indicate that mAP50-95, mAP50, and mAP75 are maintained from pixel value range of 256 to 64. In fact, some mAP values are better in PVR128 and PVR 64, compared to PVR256. However, the mAP values start to fall, starting from PVR32. Even worse, the drop in mAP becomes more drastic after PVR16. The fall in detection accuracy implies that pixel value range should be limited from 256 to 32. Otherwise, the accuracy cannot be kept in place. Comparing PVR256 with PVR32, for entropy benchmark, the average entropy decreased by about 40.6%. Looking at mAP values, mAP50-95 decreased by 0.0045%, mAP50 decreased by 0.0031%, and mAP75 decreased by 0.0062%. The values present that the accuracy barely decreased, denoting that detection accuracy is preserved and entropy of images are decreased, within the pixel value range of 256 to 32.

Through the plots of mAP50-95 value per epoch for each image processing procedure, the graph showing the growth of mAP value changing across the increase in epoch is achieved, as shown in Figure 10. Similarly to the mAP50-95 results presented in Table I, the accuracy is maintained until the point where the pixel value range is set to 32. Afterwards, mAP50-95 keeps dropping as discussed earlier. The graph shows a dent around 1 to 5 epochs for every type of image processing listed. This is due to the training process of fine tuning, which resets the trained parameters in the pre-trained model. Therefore, a hole is created within the progress in mAP value across the epochs. Otherwise, all the image processing methods follow the same shape of graph, slowly and steadily increasing and going through over-fitting in the end, except for PVR8 and PVR4.

TABLE I
Detection Accuracy and Entropy for Each Process

Process	Entropy	mAP50-95	mAP50	mAP75
PVR256	7.216	0.445	0.637	0.484
PVR128	6.232	0.447	0.640	0.485
PVR64	5.251	0.447	0.639	0.484
PVR32	4.285	0.443	0.635	0.481
PVR16	3.346	0.438	0.630	0.474
PVR8	2.448	0.423	0.614	0.456
PVR4	1.594	0.385	0.573	0.411
Sobel	6.103	0.369	0.543	0.399
Canny	0.617	0.312	0.469	0.335
LoG	3.985	0.337	0.503	0.361

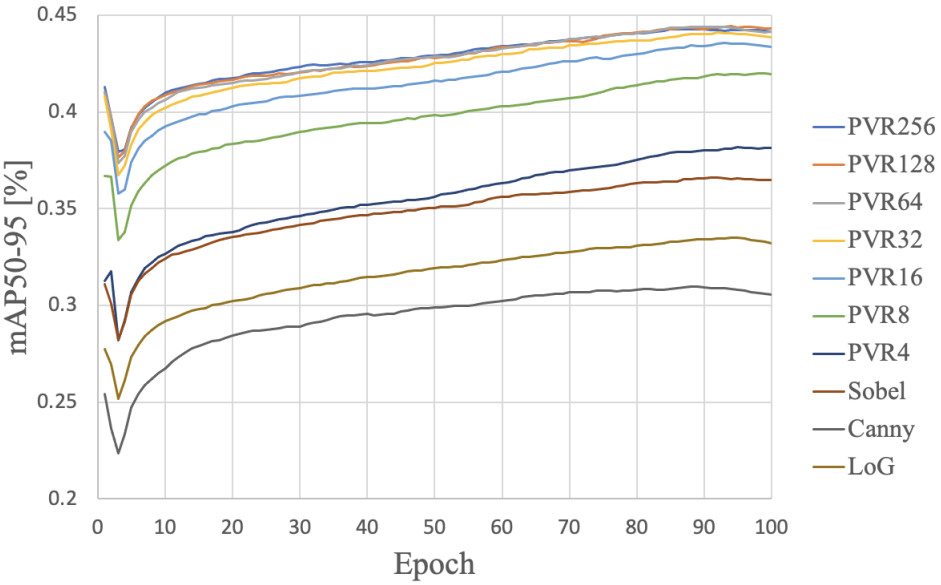


Figure 10. The relationship between the YOLOv5 mAP50-95 and epoch.

Chapter 5

Conclusion and Future Works

5.1 Conclusion

Contrast reduction method proposed in this thesis has been proven to be effective after analyzing and comparing the results with given benchmarks. The mAP values outputted by YOLOv5 are successfully maintained while also dropping the entropy as long as the pixel value range is within 256 to 32. It can be concluded that contrast reduction may be a useful technique to convert images that can be more fitting for computer vision.

5.2 Future Works

While the objective of this research has been fulfilled, there are many portions that can be improved or renewed in this thesis. The experiment has been tested on RGB images but not for YUV images. YUV is a color space used for video purposes. This implies that if YUV images behave like RGB images in terms of contrast reduction, limitation of pixel values can be applied both for image and video processing computer vision tasks. Since there is a chance that the method's application may increase, the YUV color space is worth testing out. In addition, if this method of contrast reduction can be implemented into the jpeg format, it will also work as a form of image coding. As a result, there are still several things to investigate for this research, which can be done in the future.

Bibliography

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010. DOI: 10.1109/TPAMI.2009.167.

[2] S. Jha, C. Seo, E. Yang, and G. P. Joshi, "Real time object detection and tracking system for video surveillance system," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3981–3996, 2021, ISSN: 1573-7721. DOI: 10.1007/s11042-020-09749-x.

[3] K.-K. Sung and T. Poggio, "Example-Based Learning for View-Based Human Face Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 39–51, 1998. DOI: 10.1109/34.655648.

[4] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A Survey on 3D Object Detection Methods for Autonomous Driving Applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019. DOI: 10.1109/TITS.2019.2892405.

[5] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep Learning for Medical Anomaly Detection – A Survey," vol. 54, no. 7, 2021, ISSN: 0360-0300. DOI: 10.1145/3464423.

[6] B. Xu, W. Wang, G. Falzon, *et al.*, "Livestock classification and counting in quadcopter aerial images using Mask R-CNN," *International Journal of Remote Sensing*, vol. 41, no. 21, pp. 8121–8142, 2020. DOI: 10.1080/01431161.2020.1734245.

[7] H. Choi and I. V. Bajić, "Scalable Image Coding for Humans and Machines," *IEEE Transactions on Image Processing*, vol. 31, pp. 2739–2754, Mar. 2022. DOI: 10.1109/TIP.2022.3160602.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014. DOI: 10.1109/CVPR.2014.81.

[9] R. Girshick, "Fast R-CNN," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015. DOI: 10.1109/ICCV.2015.169.

[10] W. Liu, D. Anguelov, D. Erhan, *et al.*, "SSD: Single Shot MultiBox Detector," *Computer Vision – ECCV 2016*, pp. 21–37, 2016. DOI: 10.1007/978-3-319-46448-0_2.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016. DOI: 10.1109/CVPR.2016.91.

[12] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an Image Edge Detection Filter Using the Sobel Operator," *IEEE Journal of solid-state circuits*, vol. 23, no. 2, pp. 358–367, 1988.

[13] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986. DOI: 10.1109/TPAMI.1986.4767851.

[14] R. C. Gonzales, R. E. Woods, and S. L. Eddins, *Digital image processing using MATLAB*. Pearson Prentice Hall, 2004.

[15] J. Redmon and A. Farhadi, "Yolov3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[16] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, 2017. DOI: 10.1109/CVPR.2017.690.

[17] G. Jocher, A. Chaurasia, A. Stoken, *et al.*, *ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation*, version v7.0, Nov. 2022. DOI: 10.5281/zenodo.7347926.

[18] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft COCO: Common Objects in Context," *Computer Vision – ECCV 2014*, vol. 8693, pp. 740–755, Sep. 2014. DOI: 10.1007/978-3-319-10602-1_48.

[19] J. Sachs, *Digital Image Basics*. Digital Lights and Color, 1996-1999.

[20] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

List of Figures

List of Tables

Acknowledgement

I would like to express my sincere gratitude to Professor Hiroshi Watanabe of Waseda University for his guidance on my research topic and graduation thesis, my senior colleagues Shindo, Watanabe, Iino, and Adachi for their advice and comments on my thesis writing and research, and my lab fellows Sugiyama, Nakayama, Fukuta, Ono, and Aoki for being kind and friendly, encouraging, and enlightening me.

I want to thank my teachers and friends who helped me out in daily life and gave me happiness to endure and overcome the struggles I faced.

Finally, I want to thank my family for raising me and giving me a handful help to become who I am today.