Lucas Romanó Santos

MODERN CNN-BASED OBJECT DETECTORS SYSTEMS APPLIED TO PERSON DETECTION

(*pre-defense version, compiled at December 6, 2018*)

Document presented as a partial requirement for the degree of Bachelor of Computer Science, Exact Sciences Sector, of the Federal University of Paraná, Brazil.

Field: *Computer Science*.

Advisor: David Menotti.

Curitiba PR - Brazil
2018

Resumo

A detecção de objetos passa por um momento muito promissor graças ao uso de redes neurais convolucionais (CNNs). Diversos detectores de sucesso foram propostos nos últimos anos e cada um conta com características muito diferentes em suas arquiteturas, tornando a escolha do detector de melhor desempenho para determinada aplicação uma tarefa difícil.

Escolhemos quatro arquiteturas de detectores baseados em CNN de suma importância dos últimos anos - Faster-RCNN (Ren et al., 2015), Faster-RCNN + FPN (Lin et al., 2017), RetinaNet (Lin et al., 2018) e YOLOv3 (Redmon e Farhadi, 2018) - e analisamos o desempenho de cada um no complexo e importante cenário de detecção de pessoas. Além disso, investigamos este cenário através da detecção independente de 3 diferentes áreas do corpo humano: cabeça, cabeça-ombro e corpo completo. Este trabalho se propõe como um guia para seleção de um detector para cenários similares e da investigação de casos em que pode ser possível a construção de um sistema detector mais poderoso através da união de múltiplos especializados em diferentes partes sobrepostas de um objeto de estudo. Para treino e teste dos modelos, utilizamos a base INRIA Person Dataset com novas anotações de cabeça e cabeça-ombro.

Palavras-chave: detecção de pessoas, rede neural convolucional, deep learning.

Abstract

Object detection currently goes through a very promising moment thankfully to the use of convolutional neural networks (CNNs). Several successful architectures were proposed in the last few years and each one counts with very different characteristics in its architectures, making the decision of choosing a detector for a certain application a hard task.

We choose four CNN based detectors of major importance in the last few years - Faster-RCNN (Ren et al., 2015), Faster-RCNN + FPN (Lin et al., 2017), RetinaNet (Lin et al., 2018) and YOLOv3 (Redmon e Farhadi, 2018) - and analyze the performance of each in the important scenario of person detection. Besides that, we investigate the person detection task through the independent detection of three different areas of the human body: head, head plus shoulder and full body. This work is proposed as a guide for selection of a detector for similar scenarios and the investigation of cases that it's possible to build a more powerful detector systems through the combination of multiple detectors specialized in different overlapping regions of an object of study. For training and testing the models, we use the established INRIA Person Dataset with new annotations for head and head plus shoulder regions.

Keywords: pedestrian detection, convolutional neural network, deep learning.

List of Figures

List of Tables

Contents

1 Introduction

The task of object detection is an area of study of computer vision which goal is to identify objects in a image or video. In the last years, the proposal of deep learning models based on Convolutions Neural Networks (CNNs) drastically improved the accuracy and efficiency, making this task feasible for various industries, including in mobile applications.

Common images and videos primarily focus on people: about 35% of pixels in movies and YouTube videos as well as about 25% of pixels in photographs belong to people (Laptev, 2013). Person detection attracted much attention in the last years due to development of self-driving cars, surveillance and robotics. However, it is a very challenging problem because different scenario conditions, luminosity, angle and scale of a person produce very sparse features. As detectors have different meta-architectures and feature-extractors, it is clear to say that each one behaves very differently for the same input.

Full-body human detection often suffers scenes in which people are not standing and from occlusions among individuals. Besides that, many applications do not attempt to detect a person head to toe, like face recognition and human tracking. Hence, much research emerged then focused in the upper part of the human body, specifically in the head and head plus shoulder region. We address each type of region as *head* and *head-shoulder* respectively, and *body* when referring to the entire person. Figure 1.1 compares each region of interest.

The detection of each one of the described regions of interest resulted in a number of different methods and custom architectures within the scientific community, usually related to a particular case of use. *Head* and *head-shoulder* detection related publications are usually related to people counting, while *body* detections applications usually are related to pedestrian detection, surveillance or robotics. Regarding CNN-based detectors, it is known that different objects, even if similar for human glance, produce completely different features to be considered and consequently the behaviour and performance of each detector also differ.

In this dissertation, we select four of the most relevant and promising modern object detectors - Faster-RCNN (Ren et al., 2015), Faster-RCNN + FPN (Lin et al., 2017), RetinaNet (Lin et al., 2018) and YOLOv3 (Redmon e Farhadi, 2018) - and apply them in the scenario of person detection. We present the performance of each selected model for each type of labeling and compare its accuracy and speed. Besides, we compare how it's possible to combine the resulting inferences of each detector regarding a region of interest, detecting more persons. We show that YOLOv3, as at least 3x faster than all others state of art detectors, can produce much

Figure 1.1: Comparison of regions of interest. We address each type of region as *head*, *head-shoulder* and *body*, respectively.

more accurate results when combining inferences aimed to different regions of interest (*head*, *head-shoulder* and *body*) against other more accurate detectors, while still remaining the fastest.

1.1 Contributions

We summarize our contributions as two:

- We provide a concise survey of modern CNN-based detectors and analyze the performance of each in the scenario of person detection.
- We explore the task of person detection considering multiple overlapping regions of interest (head, head-shoulder and body) and analyze how to amount of features of these regions contributes for the accuracy and speed for each analyzed detector. We hope that the this type of implementation may be useful to cases that detecting the entire bounding box of an object is not a requirement or confidence of detections needs to be improved.

1.2 Document Structure

The remainder of the dissertation is organized as follows. In Chapter 2 we highlight some of the most relevant related publications. Chapter 3 introduces the extended INRIA Person Dataset, used for experimenting in this dissertation. An overview of the analyzed detectors is presented in Chapter 4, followed by considerations of the detectors and training in Chapter 5. Experimental results and analysis are presented in Chapter 6. Finally, we draw our conclusion in Chapter 7.

2 Related Work

In this chapter, we present a review of the most relevant studies related to this dissertation. We start by presenting a historic approach for publications involving the general object detection, followed by person detection studies. We then present specialized detectors created specifically for the task of head detection and head plus shoulder detection. Finally, we describe related work referring to the comparison of object detectors.

2.1 Object Detection

The earliest successes object detectors were based on the **sliding window** paradigm. The image is divided into several windows and it's attributed for each one the score of a potential detection using the content inside that window. LeCun et al. (1989) applied convolution neural networks to the recognition of handwritten zip codes. Viola Jones face detection framework (Viola e Jones, 2001) became one of the biggest marks in the area being the first with real-time detector with competitive rates, later applied to many others objects of study.

The **two-stage detector** paradigm is divided in two parts: the first is responsible for generating proposals and the second is responsible for analyzing these proposals into foreground classes or background. R-CNN (Girshick et al., 2016) method applied a convolutional neural network to the second stage and obtaining impressive results creating a new era in object detection. R-CNN was improved with the introduction of SSPnet (He et al., 2014), drastically improving the detector efficiency by sharing computation through a feature pyramid which is robust to region size and scale. However, training was very complex and expensive because both R-CNN and SSPnet required a complex multi-stage pipeline. Fast R-CNN (Girshick, 2015) proposed a solution to this with a single-stage training, using a multi-task loss, while still improving accuracy.

Faster R-CNN (Ren et al., 2015) introduces the Region Proposed Networks (RPNs), integrating the proposal generation stage with the second-stage classifier into a single convolutional network obtaining nearly cost free region proposal generation. Several improvements to this detector were proposed, like (Shrivastava et al., 2016), (Lin et al., 2017), (He et al., 2016) and more recently Mask R-CNN (He et al., 2017). Mask R-CNN efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance.

We stick to experimenting Faster R-CNN detector since Mask R-CNN is focused on instance segmentation and demands segmentation labeling, which is not the focus of this dissertation.

In parallel, **one-stage** paradigms were explored and YOLO (Redmon et al., 2016), SSD (Liu et al., 2016) and later RetinaNet (Lin et al., 2018) emerged. This paradigm refers to architectures based on a single feed-forward convolutional network that predict classes and it's confidence without requiring a second stage classification operation. In this paradigm it's observed that usually accuracy is sacrificed in trade of speed, with YOLO (Redmon et al., 2016) pushing this boundary even more.

YOLO was improved to YOLOv2 (Redmon e Farhadi, 2017) and severely improved its accuracy while still remaining the fastest, although its accuracy was still lower than SSD and much lower than Faster R-CNN. RetinaNet, later introduced, provides both excellent accuracy and speed. The subsequent version of YOLOv2, YOLOv3 (Redmon e Farhadi, 2018), does not reach RetinaNet accuracy but keep its position as the fastest until the present.

2.2 Person Detection

Person detection as one of the main cases in the area of object detection has a rich and continuous history. Classic methods commonly use haar-like wavelets (Zhang et al., 2014) or histograms of gradient orientation (HOG) (Dalal e Triggs, 2005) as features, and support vector machine (SVM) (Alonso et al., 2007) or AdaBoost (Stauffer e Grimson, 2000) as the verification methods on top of a sliding window based paradigm for proposal generation. At the time, the best classical architectures lead to very similar performance. (Benenson et al., 2014).

More recently introduced, pedestrian detectors like (Tian et al., 2015a) and (Tian et al., 2015b) are hybrid methods that combines traditional hand-crafted features and deep convolutional features. Although generic object detectors are successful in person detection, the results are not optimal as hand-crafted methods has been shown to be of critical importance for the task of pedestrian detection: Zhang et al. (2016) demonstrates that Faster R-CNN alone have limited success in the task of pedestrian detection and propose a detector based on RPN and custom features optimized for pedestrian detection.

More recently, SAF R-CNN (Li et al., 2018) was build on top of Faster-RCNN proposes a unification of a large-size sub-network and a small-size sub-network, reaching state-of-the-art performance. The proposal of these type of detectors aims to reach top performance in the task of person detection. However, the goal of this dissertation is to compare the performance of general object detectors in this complex scenario.

Most of the head or head-shoulder region detection approach were addressed specifically to the problem of people counting due to the condition of recurrent partial occlusion. At 2.2.1 and 2.2.2 we highlight detectors depending on its region of interest.

2.2.1 Head Detection

The work of Lin et al. (2001), one of the pioneers of head detection, applied Haar wavelet transforms to extract feature areas of head-like contours to be analyzed by a support vector machine (SVM) that would classify the area as head or not.

Although the task of face detection reached considerable maturity (Mathias et al., 2014), the more general task of head detection remained challenging. With the advent of deep learning, several new methods were proposed in the last few years for this task. Gao et al. (2016) applies the AdaBoost algorithm to obtain the head region proposals, which are used as input for a CNN to produce head region proposals that will finally be classified by a linear support vector machine (SVM).

More recently, Vu et al. (2015) extended the R-CNN object detector (Girshick et al., 2014) combining multiple CNNs into a joint framework. Crowdnet (Boominathan et al., 2016) combines deep and shallow fully convolutional networks to predict the density map for a given crowd image through target the head region.

2.2.2 Head-shoulder Detection

Kilambi et al. (2008) proposed a blob-based method to estimate the number of people in urban environments. However, blob-based methods can only detect moving objects and shadows greatly affect the total count.

Li et al. (2008) applied the Histograms of Oriented Gradients (HOG) feature (Dalal e Triggs, 2005) to detect the head-shoulder region. However, the method requires several seconds to process a 320x240 image, which was too slow for real application. An extension of this was proposed by Zeng e Ma (2010a) greatly improving the performance with a novel method combining a Viola-Jones type classifier and a HOG feature based AdaBoost classifier, followed by a particle filter tracker using local HOG features. Zeng e Ma (2010b) combines HOG with LBP (Local Binary Pattern) reaching even more performance. We note that these proposers are stand-alone pedestrian detectors consisting of hand-crafted features and boosted classifiers.

In the last few years, no more outstanding publications were proposed when targeting head-shoulder regions. We also highlight that no head-shoulder CNN-based detector publications were found.

2.3 Comparison of Object Detectors

Canziani et al. (2016) compares multiple state-of-the-art deep neural networks submitted to the ImageNet challenge (Russakovsky et al., 2015), providing insights into the design choices that can lead to efficient neural networks for practical application. Huang et al. (2017) explores the speed, accuracy and memory trade-offs of modern detection systems and the influence of

characteristics like feature extractor and image resolution through implementations of the Faster R-CNN, R-FCN and SSD detectors in Tensorflow.

Both analysis are made in a very low-level, exploring deeply the performance details of each studied architecture. Although such comparisons may be very useful to optimizing a classification system, this is not the focus of this work. We focus on providing a high-level overview of modern detectors systems to person detection and study the combination of multiple detectors inferences aiming to different overlapping regions.

3 Dataset

For training and evaluating the detectors and their results presented later in this dissertation we use the established INRIA Person Dataset [1].

The dataset consists of images of humans taken from several viewpoints under varying lighting conditions, ranges, human poses and cases of occlusion in both indoor and outdoor activities. Image 3.1 contain some examples.

Figure 3.1: Examples of INRIA Person Dataset.

The dataset includes also the cropped version of the images. Unlike the original authors we disconsider them and use only the full images for both training and testing the detectors. We do so, in order not only to evaluate the detector in terms of false positive detections, but because the analyzed detectors architectures do consider the background of a image and training it would actually be prejudicial to the training. Besides, this gives a more realistic assessment on how well a detector performs for real world applications. The negatives images are also ignored since the selected detectors are trained only through positive cases. This results to 615 images for training and 288 for testing.

As explained, we consider 3 types of labeling for training: *head*, *head-shoulder* and *body*. For the body labeling type we use the original dataset annotations provided. The labels for *head* and *head-shoulder* were manually created only to persons which were already annotated by the original dataset: cases that a person is visible in the image but not originally labeled in the dataset are ignored for keeping the training fairer and the modified dataset consistent, resulting in 1235 occurrences for each label in the training dataset and 589 for the testing dataset.

The size of the object is one of the main variables to be considered. The established general object recognition dataset COCO (Lin et al., 2014), for example, categorizes the classes

[1] http://pascal.inrialpes.fr/data/human/

depending on the average size for comparison of the detectors. Generally speaking, CNN-based object detectors do perform worse for smaller objects and more details are provided in the next chapter.

Images of the dataset have variant sizes from 176x232 to 1298x976 pixels and Figure 3.2 compares the distribution of the bounding boxes sizes of all 3 labels in the modified dataset. Table 3.1 compares the same metrics regarding mean and standard deviation.

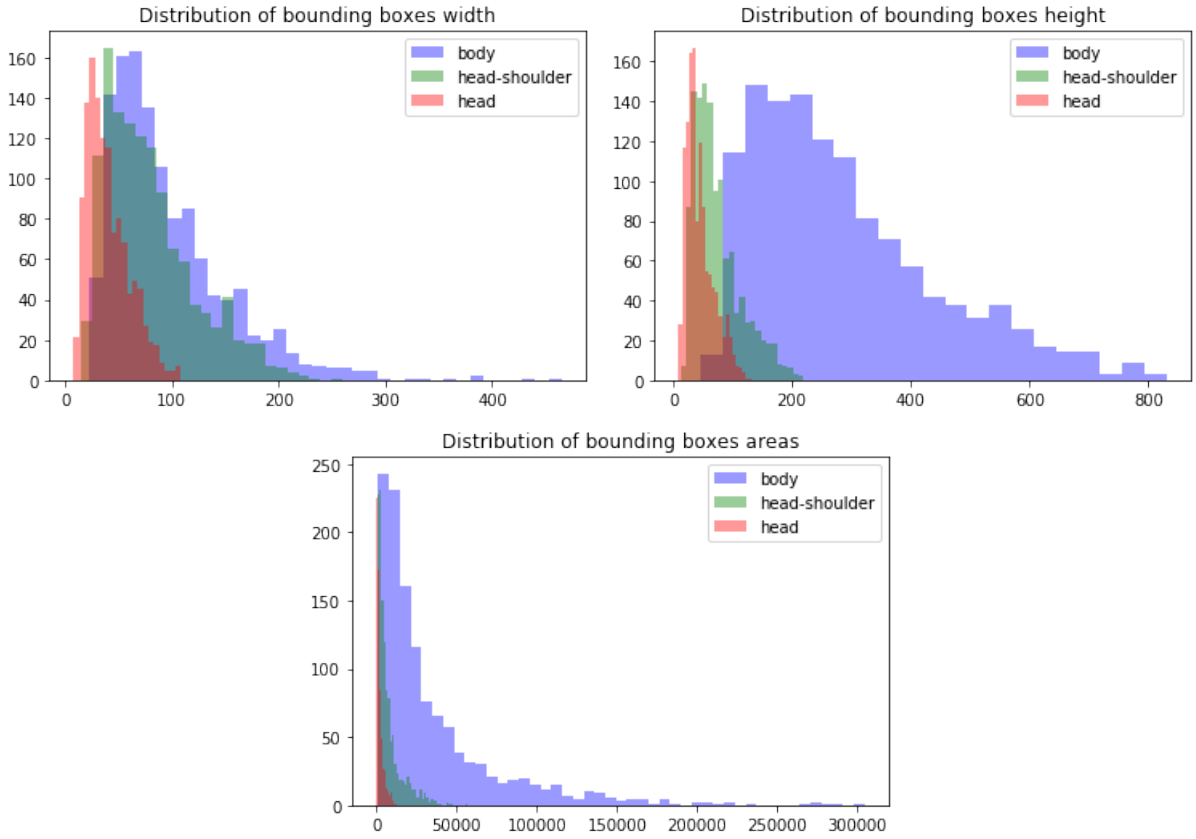

Figure 3.2: Distributions of bounding box width, height and areas of the modified INRIA Person Dataset. Better visualized in colors.

	width		**height**		**area**	
	mean	std	mean	std	mean	std
head	39.5	19.7	44.4	23.2	2201.7	2238.6
head-shoulder	79.1	42.0	72.3	39.3	7269.6	7898.7
body	96.8	56.5	289.4	156.8	36058.9	41645.3

Table 3.1: Mean and Standard Deviation for width, height and area of bounding boxes in the INRIA Person Dataset. Standard deviation is abbreviated as *std*.

The modified dataset is publicly available online [2].

[2]https://web.inf.ufpr.br/vri/

4 Detectors Overview

We selected the most relevant detectors in the last years based on its performance and influence to subsequent works. In the next sessions, we briefly present the detectors analyzed in this dissertation. Details about the implementations and training are presented in the next chapter.

4.1 Faster R-CNN and Faster R-CNN + FPN

As a two-stage detector, Faster R-CNN (Ren et al., 2015) is composed of two networks: the first is responsible for generating region proposals and the second to analyze these proposals. Instead of using Selective Search (SS) like its predecessor introduced at (Girshick, 2015), Faster R-CNN introduces Region Proposal Networks (RPNs), which is able to generate accurate and efficient proposals nearly cost-free.

Since it's creation in 2015, Faster R-CNN and it's variants (Shrivastava et al., 2016) (Lin et al., 2017) (He et al., 2016) has been very influential due to its excellent accuracy (see table 4.1) and by popularizing the concept of anchors, later used for many detectors (including YOLOv2, YOLOv3 and RetinaNet). Anchors are predefined bounding boxes that are placed through the image with different sizes and ratios and these bounding boxes act as a reference when first predicting object locations. In the case of Faster R-CNN, the RPN predicts the possibility of the proposals originated by the anchors are background or foreground and refine these anchors. The many proposals generated are then filtered by a Region of Interest (ROI) layer to generate features of the same size.

In this dissertation, we also analyze an extension of Faster R-CNN which is integrated with Feature Pyramids (FPN) (Lin et al., 2017). In brief, FPN constructs rich, multi-scale feature pyramid from a single resolution input image with a top-down pathway and lateral connections. We chose this extension since it holds best results (see table 4.1) as a two-stage detector in the PASCAL VOC metric Everingham et al. (2006), which results are provided in this dissertation.

4.2 RetinaNet

As YOLO, RetinaNet is a one-stage object detector. It is composed of a backbone network and two task-specific subnetworks. The backbone network computes a convolutional feature map over

	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods						
Faster R-CNN+++ (He et al., 2016)	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN (Lin et al., 2017)	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI (Huang et al., 2017)	34.7	55.5	36.7	36.7	13.5	38.1
Faster R-CNN w TDM (Shrivastava et al., 2016)	36.8	57.7	39.2	16.2	39.8	**52.1**
One-stage methods						
YOLOv2 (Redmon e Farhadi, 2017)	21.6	44.0	19.2	19.2	5.0	22.4
SSD513 2016 (Fu et al., 2017)	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 (Fu et al., 2017)	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet (Lin et al., 2018)	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet (Lin et al., 2018)	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	51.2
YOLOv3 608 × 608 (Redmon e Farhadi, 2018)	33.0	57.9	34.4	18.3	35.4	41.9

Table 4.1: Average Precision of general object detectors on COCO TEST-DEV. Table obtained from (Redmon e Farhadi, 2018) with the backbone information occluded.

an entire input image and is an off-the-self convolutional network. It is also an adaptation of FPN due to its efficiency to construct rich and multi-scale feature pyramids: each level of the pyramid can be used for detecting objects at a different scale. The first subnetwork performs convolutional object classification on the backbone's output, while the second performs convolutional bounding box regression.

It is pointed that the main reason that single-stage detectors reach lower accuracy is the extreme foreground-background class imbalance encountered during training of dense detectors is the central cause. These detectors evaluate up to 10^5 candidates locations, but only very few have objects. This leads to an inefficient training as most locations are easy negatives that do not contribute to learning and the large quantity can overwhelm training and lead to degenerate models. Instead of downsampling dominant or oversampling minority cases like Faster R-CNN, the publication proposes a new loss function called *Focal loss* for classification, which significantly increased the accuracy.

RetinatNet currently holds state of art performance as most accurate detector on COCO TEST-DEV, surpassing even two-stage detectors. Table 4.1 displays the average precision of multiple modern detectors on COCO TEST-DEV. Besides, provides excellent trade-off with its speed, with about 70ms inference time on it's fastest architecture. See figure 6.2 for details.

4.3 YOLOv3

First introduced in Redmon et al. (2016), the YOLO detector as a one-stage detector is able to direct predict bounding boxes and class probabilities with a single network in a single evaluation. Generally, the system divides the input image into a grid and each grid cell is responsible for detecting the object if the center of the object is inside it, predicting bounding boxes and a confidence score for each. Since each grid cell outputs many bounding boxes, a Non Maximum Suppression algorithm (NMS) is responsible for merging the bounding boxes of a same object into one.

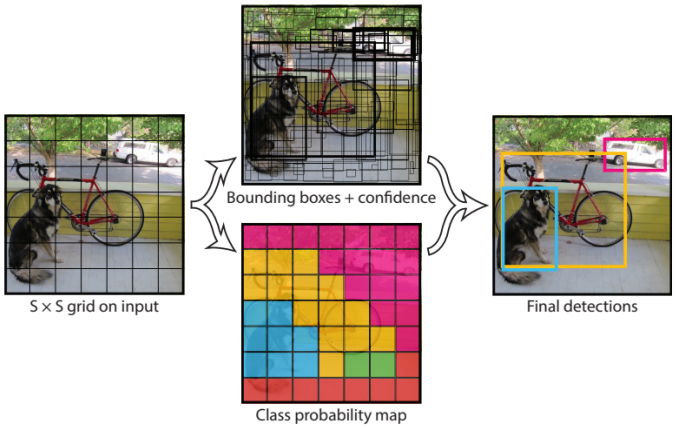

Figure 4.1: Representation of the YOLO Detector. Image obtained from Redmon et al. (2016).

YOLO makes a significant number of localization errors and has relatively low recall compared to region proposal-based methods and improvements to these were presented in YOLOv2 (Redmon e Farhadi, 2017). Higher resolution images are accepted as input (the images are resized randomly to different resolutions): the YOLO model uses up to 448x448 images while the YOLOv2 uses up to 608x608 images, improving the detection of smaller objects. Besides, batch normalization is added to prevent overfitting and the final fully connected layers are modified to the use of anchor boxes to help prediction of the bounding boxes.

More improvements are followed in YOLOv3 (Redmon e Farhadi, 2018). The most salient feature is that the detection happens in 3 different layers with different scales, greatly improving the detection of smaller objects, the main shortcoming of YOLOv2. Although it does not reach top accuracy (see table 4.1), YOLOv3 still reaches competitive results and is by far the fastest detector, with inference times about 22ms on its fastest architecture. See figure 4.2 for details.

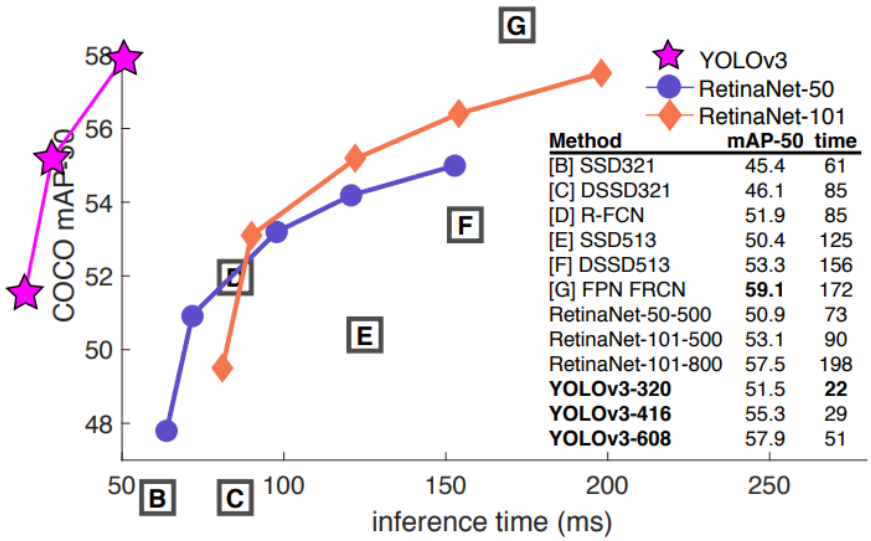

Figure 4.2: Speed (ms) versus accuracy (AP) on COCO TEST-DEV. Image obtained from Redmon e Farhadi (2018).

5 Detector Considerations

The INRIA Person dataset has a considerable small number of images compared to new datasets available in the scientific community and to keep compatibility with real world problems, we keep standard data augmentation in training for all detection systems. Since the purpose of this dissertation is to provide on overview of the performance of the selected detectors in the scenario of person detection, no fine-tuning is performed, and we do not try to reach peak performance of each detector. The results shown here are proposed as demonstration and reference to real world applications. In brief, training is made simply with the selection of the optimal model based on the validation loss through iterations in training and no longer than 10 hours.

Below we present relevant information about the evaluated detector systems:

- Faster-RCNN and Faster-RCNN + FPN: we use the Detectron implementation (Girshick et al., 2018) and the system is trained with end-to-end training (Ren et al., 2015) since we desire to train the entire network as a single training using all four loss function (rpn regression loss, rpn objectness loss, detector regression loss and detector class loss).

- RetinaNet: we use the Detectron implementation (Girshick et al., 2018). The chosen base model is ResNet-101-FPN backbone, which is presented with the best results in (Lin et al., 2018) for the COCO challenge.

- YOLOv3: Our initial goal was to use only Keras-based implementations for all detectors, but it was discontinued due to the facilities provided by Detectron. We use an open-source implementation available at (qqwweee, 2018). We highlight that this detector is a custom implementation of the original detector and the results may slightly differ from what would be with the original detector. This detector is addressed as YOLOv3 for simplicity.

All models were initialized with pretrained weights for ImageNet classification (Russakovsky et al., 2015), provided by the respective repository authors. More details of training and analysis can be found at [1].

All presented experiments were executed in a NVIDIA Titan X provided by the Laboratory of Vision, Robotics and Imaging (VRI), part of the Department of Informatics in the Universidade Federal do Paraná.

[1] https://github.com/lucasromanosantos/person-detection-comparison

6 Result

In this chapter we report our results and comparisons of the selected detectors in the conditions described in Section 5. All results here presented were experimented on the modified INRIA Person Dataset TEST. We divide this section regarding the variable of interest of the analysis.

6.1 Effectiveness

We use the average precision (AP) to evaluate the accuracy of the detector. Ground truth and final detections are matched using the PASCAL criterion (Everingham et al., 2006), which demands a minimum overlap of 50% for two matching bounding boxes. The average precision is defined as $\sum_{i=0}^{N} P_i \times \triangle R_i$ where N is the total number of images in the collection, P_i is precision at a cutoff of i images, and $\triangle R_i$ is the subtractions of recall between cutoff $i-1$ and cutoff i. In practice, this is closely approximated to the area under the precision/recall curve. Precision and recall are defined as:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Table 6.1 shows the AP for each detector. Generally we observe that the *head* labeling, which bounding boxes are smaller and provide fewer features, consistently loses accuracy in comparison to the others. The *head-shoulder* label results are very promising, with even higher accuracy on YOLOv3 and RetinaNet than *body*.

AP	Faster R-CNN	Faster R-CNN + FPN	RetinaNet	YOLOv3
	(2015)	(2017)	(2018)	(2018)
head	0.8613	0.8805	0.6601	0.7047
head-shoulder	0.9081	0.8869	0.8328	0.8712
body	0.9203	0.9191	0.8317	0.8002

Table 6.1: Average Precision of each detector for each region of interest. We also highlight the year of publication of each detector.

Surprisingly, the oldest detector (Faster R-CNN) had the best AP of all detectors for all regions, barely surpassing it's improved version (Faster R-CNNN + FPN). Intuitively, stronger

	Faster R-CNN		**Faster R-CNN + FPN**		**RetinaNet**		**YOLOv3**	
	TP	FP	TP	FP	TP	FP	TP	FP
head	517	58	520	29	390	2	470	143
head-shoulder	538	35	523	11	491	8	518	31
body	544	33	542	15	490	3	473	6

Table 6.2: Relation of True and False Positives of each detector for each region of interest.

performance on classification should be positively correlated with stronger performance on COCO TEST-DEV detection, presented in Section 4. In this case, this was not true: although FPN is part of architecture of the others three detectors (in the case of YOLOv3, a similar structure), Faster R-CNN still reached better performance. We also highlight the poor performance of YOLOv3 and RetinaNet for the *head* labeling. On the other side, both detectors performed better with a medium amount of features, with higher AP on *head-shoulder* than *body* detection.

Table 6.2 compares the True Positives (TP) and False Positives (FP) inferences of each detector. Except for RetinaNet, we observe the *head* label suffers with an alarming FP rate, drastically bigger for YOLOv3. Generally speaking, *head-shoulder* and *body* labeling is much more stable, with *body* having lesser cases of false positives. RetinaNet has shown to be very sensible for avoiding false positives, with less occurrences for all cases. Although its accuracy was not as satisfactory, YOLOv3 FPs occurrences for *body* is drastically lower than head-shoulder, which counts with 7% more AP.

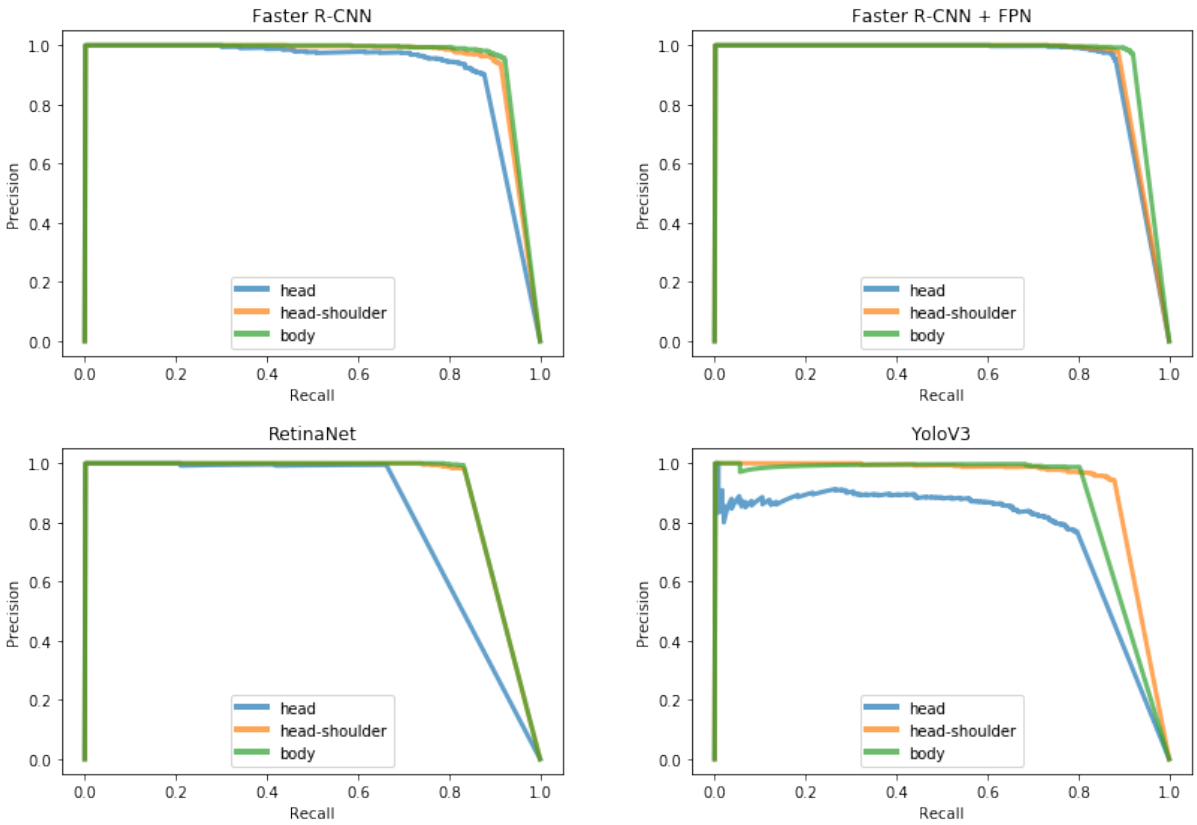

Figure 6.1: Precision vs Recall curve of *head*, *head-shoulder* and *shoulder* for each detector.

In figure 6.1 we display the precision vs recall curve categorized by region for each detector. The precision/recall curve is calculated with precision monotonically decreasing by setting the precision for recall r to the maximum precision obtained for any recall $r' > r$. All detectors have a very high performance and similar curves, except for YOLOv3, as expected due to the many occurrences of false positives for the *head* label. We also observe that despite the better AP of Faster R-CNN, the FPN version significantly has a stabler precision/recall curve due to its lower amounts of FPs.

6.2 Speed

Figure 6.2 shows the average inference time of each detector. The speed of each very similar to what is reported in the respective publications, with the exception of RetinaNet with generally higher inference times than Faster R-CNN + FPN.

The first thing that is very apparent is that despite being the most accurate, Faster R-CNN is by far the slowest of all, with inference times approximately 10 times bigger than YOLOv3 and 4 times it's improved version with FPN. It is expected that RetinaNet to be faster than Faster R-CNN for being a single-stage detector and due to comparison in its original publication. In our experiments this was not the case, with Faster R-CNN + FPN being 1.5 faster. Further investigation is needed regarding this behavior.

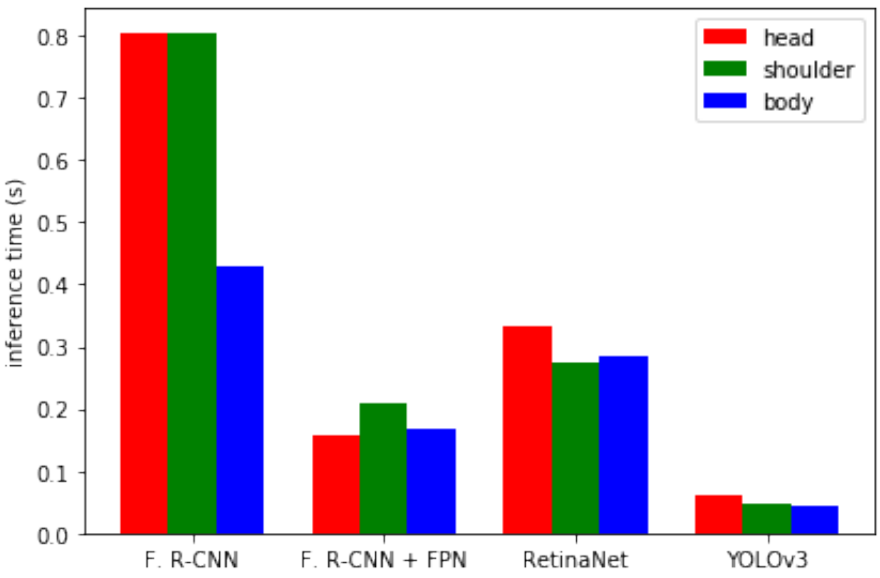

Figure 6.2: Inference times of each detector for each region of interest. We remove the inference time of the first image as caches and auto-tuning need to warm up.

6.3 Combining inferences

This section investigates the joint of resulting bounding boxes of multiple specialized detectors. This approach is much more computationally feasible nowadays due to the increasing speed of CNN-based detectors, enabling the creation of a more powerful detector when more accuracy or confidence of inferences is desired. Figure 6.3 displays an example of the same input image to detectors trained in the different regions of interest.

Figure 6.3: Example of inferences to the same image. The figure displays inferences of RetinaNet to *head*, *head-shoulder* and *body*, respectively. Combining the inferences would result in the correct detection of three pedestrians.

Figure 6.4 compares the percentage of detected persons. These results are provided with the removal of all false positives occurrences, thus we only evaluate recall improvements of the combined detectors. As false positives are disconsidered, it's important to keep in mind that this approach benefit the presented high false positives detectors and do not correspond to out of the box results.

The first thing that is very apparent is that the combination of all 3 detectors do not provide outstanding recall results compared to the combination of only two regions. In the case of RetinaNet, for example, the recall almost the same as the combination of only *head-shoulder* and *body* regions. For the others detectors, the benefits are minimal.

In cases where the AP of certain region is low, the benefits of combining it with other is also very low. RetinaNet counts with 66.21% *head* recall and combining it with *head-shoulder* or *body* resulted in increasing recall of less than 2%. YOLOv3 would suffer from a similar behavior but in fewer proportions due to the removal of false positives.

As expected, the benefits of combining the regions are very corresponding to the AP of each independent region, but not strictly attached to it. In the case of Faster R-CNN, we highlight that combining *head* plus *head-shoulder* and *head-shoulder* plus *body* resulted in the same recall, even that the AP for head is 4% lower than *head-shoulder*. In the case of Faster R-CNN + FPN, *head* plus *body* was the only one to surpass *head-shoulder* plus *body* combination and surprisingly equalizing the recall of the combination of all 3 regions.

This approach has been shown very beneficial for faster and with lower AP detectors. YOLOv3 has been able to reach 93.55% recall and still would be faster than all the others

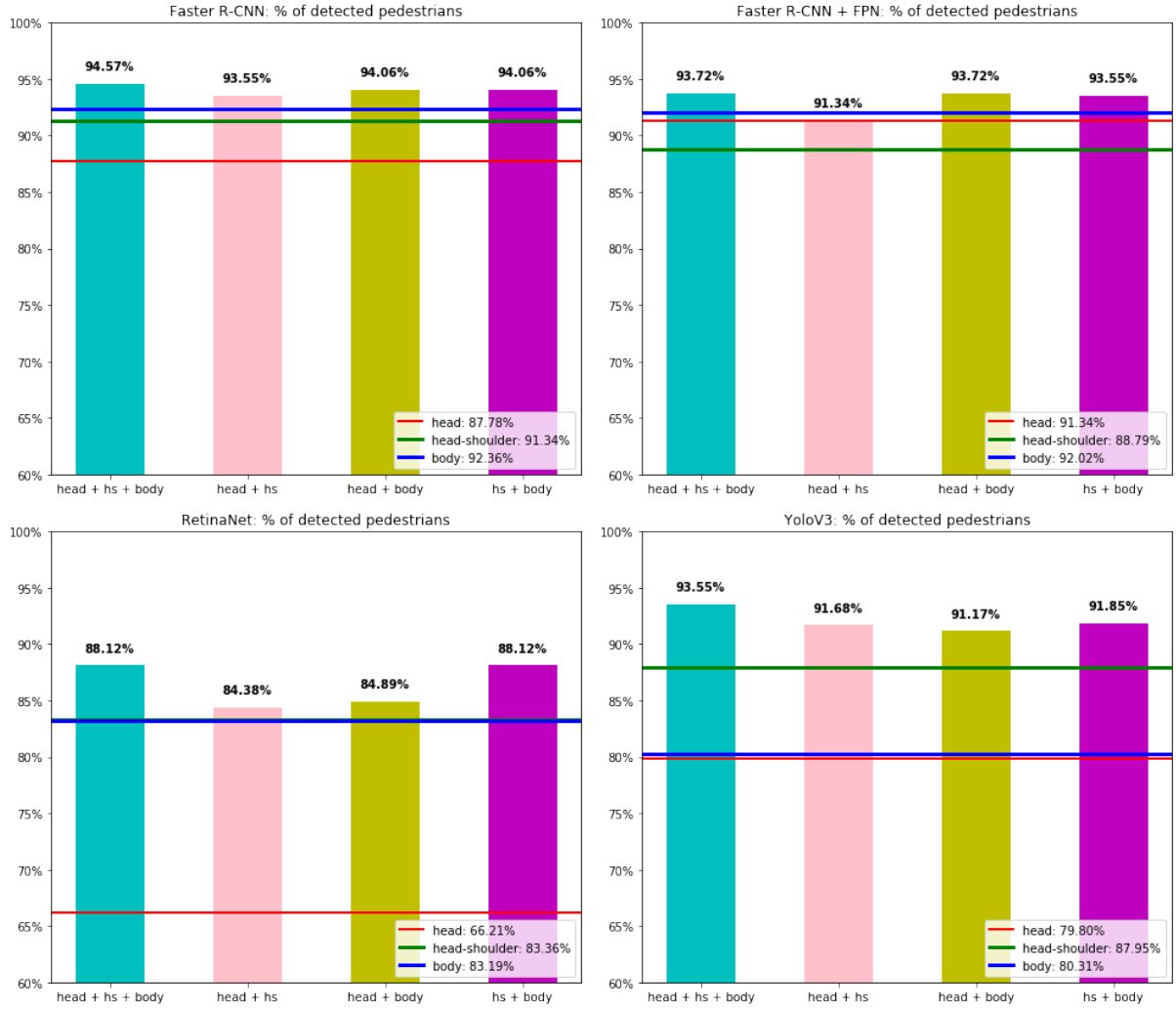

Figure 6.4: Gains of accuracy when combining the specialized detectors. The horizontal lines display the accuracy of the individual respective detector. False positives are disconsidered.

detectors. Comparing to the *head-shoulder* only detection, the increase of recall would be almost 6%. *Head* and *body* regions combination is even more outstanding, with at least 10% of recall increase. However, further investigation is required on false positives cases and the resulting AP.

In the scenario of person detection, it is clear to say that combining the detection of two regions is very beneficial. The benefits are more outstanding if the detector has a lower AP for the regions, which is expected of YOLOv3. However, YOLOv3 is much faster, making it much more computationally feasible. If computational power is available and the possibility of multiple labeling of the images in the dataset, combining the results should be considered if more recall or increased confidence is desired. Figure 6.5 display the inferences and the corresponding confidences of all detectors for the same input image.

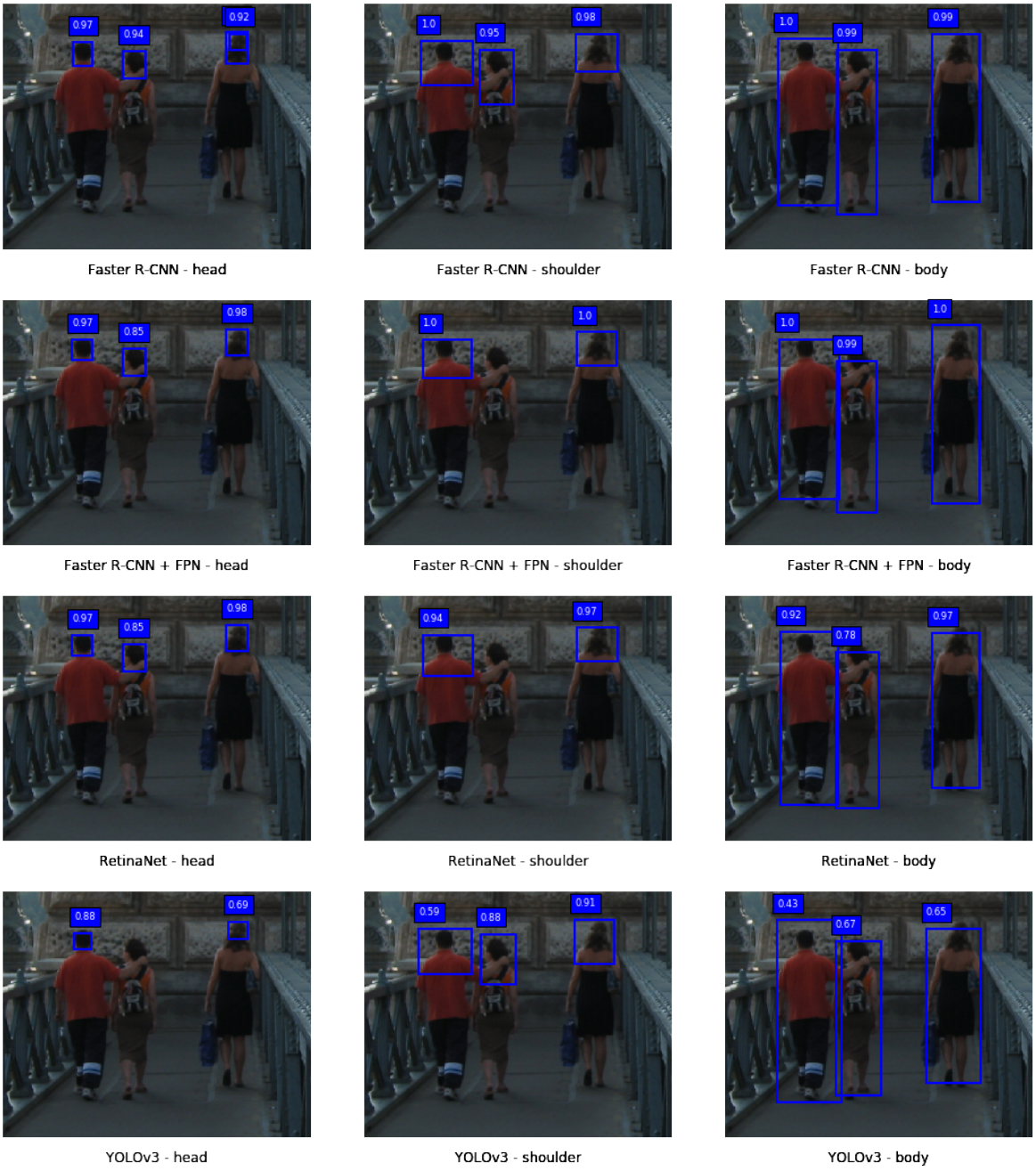

Figure 6.5: Comparison of inferences between different detectors and regions of interest.

7 Conclusion and Future Work

In this dissertation we applied and explored results of multiple state-of-art CNN-based general object detectors to person detection. Our goal is to provide a general performance of each detector that can lead to the choice of the best suited detector for a real-world application. We evaluated the performance of each in the task of person detection and demonstrated that although RetinaNet holds the best results for the general task of object detection, further analysis of variables like number of classes and training dataset size is required to the choice of the best suited detector. For the presented conditions, Faster R-CNN was shown to be the most accurate, even being the oldest of all.

Regarding person detection, we demonstrated that both detection aimed to the head or head plus shoulder regions shows competitive results compared to aiming the complete body and should be considered in this scenario. We also explored the detection of multiple regions of interest, applying different labeling of overlapping regions (*head*, *head-shoulder* and *body*) and explored the performance of each, including the possibility of combining the results of each detector. This may be an option not only for person detection but for other complex objects detections when more accuracy or confidence of inferences is desired. For this task, we concluded that this approach is more beneficial when the average precision is lower and the detector faster, which was the ideal case for YOLOv3. However, our analysis only considered the improvements of recall, requiring more research in the impact to precision.

Although the experimental results give a general idea, further improvement is required to the fairest training of the detectors: as explained, the training is made simply, with no exploration of optimal trained models. The training process includes much more variables that may work more or less depending on the architecture of the detector. The more correct or fair training greatly influences both accuracy and speed. Besides, another possible area of investigation is to study the combination of the multiple specialized detectors. While the naive approach would be to run the multiple object detectors at the same time, it's important to explore the optimal combination of these into a single CNN.

References

Alonso, I. P., Llorca, D. F., Sotelo, M. Á., Bergasa, L. M., de Toro, P. R., Nuevo, J., Ocaña, M. e Garrido, M. Á. G. (2007). Combination of feature extraction methods for svm pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):292–307.

Benenson, R., Omran, M., Hosang, J. e Schiele, B. (2014). Ten years of pedestrian detection, what have we learned? Em *European Conference on Computer Vision*, páginas 613–627. Springer.

Boominathan, L., Kruthiventi, S. S. e Babu, R. V. (2016). Crowdnet: A deep convolutional network for dense crowd counting. Em *Proceedings of the 2016 ACM on Multimedia Conference*, páginas 640–644. ACM.

Canziani, A., Paszke, A. e Culurciello, E. (2016). An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*.

Dalal, N. e Triggs, B. (2005). Histograms of oriented gradients for human detection. Em *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, páginas 886–893. IEEE.

Everingham, M., Zisserman, A., Williams, C. e Van Gool, L. (2006). The pascal visual object classes challenge 2006 (voc 2006) results.

Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A. e Berg, A. C. (2017). Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*.

Gao, C., Li, P., Zhang, Y., Liu, J. e Wang, L. (2016). People counting based on head detection combining adaboost and cnn in crowded surveillance environment. *Neurocomputing*, 208:108–116.

Girshick, R. (2015). Fast r-cnn. Em *Proceedings of the IEEE international conference on computer vision*, páginas 1440–1448.

Girshick, R., Donahue, J., Darrell, T. e Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. Em *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 580–587.

Girshick, R., Donahue, J., Darrell, T. e Malik, J. (2016). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158.

Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P. e He, K. (2018). Detectron. `https://github.com/facebookresearch/detectron`.

He, K., Gkioxari, G., Dollár, P. e Girshick, R. (2017). Mask r-cnn. Em *Computer Vision (ICCV), 2017 IEEE International Conference on*, páginas 2980–2988. IEEE.

He, K., Zhang, X., Ren, S. e Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. Em *European conference on computer vision*, páginas 346–361. Springer.

He, K., Zhang, X., Ren, S. e Sun, J. (2016). Deep residual learning for image recognition. Em *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 770–778.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S. et al. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. Em *IEEE CVPR*, volume 4.

Kilambi, P., Ribnick, E., Joshi, A. J., Masoud, O. e Papanikolopoulos, N. (2008). Estimating pedestrian counts in groups. *Computer Vision and Image Understanding*, 110(1):43–59.

Laptev, I. (2013). *Modeling and visual recognition of human actions and interactions*. Tese de doutorado, Ecole Normale Supérieure de Paris-ENS Paris.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. e Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Li, J., Liang, X., Shen, S., Xu, T., Feng, J. e Yan, S. (2018). Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, 20(4):985–996.

Li, M., Zhang, Z., Huang, K. e Tan, T. (2008). Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. Em *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, páginas 1–4. IEEE.

Lin, S.-F., Chen, J.-Y. e Chao, H.-X. (2001). Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(6):645–654.

Lin, T.-Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B. e Belongie, S. J. (2017). Feature pyramid networks for object detection. Em *CVPR*, volume 1, página 4.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. e Dollár, P. (2018). Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. e Zitnick, C. L. (2014). Microsoft coco: Common objects in context. Em *European conference on computer vision*, páginas 740–755. Springer.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. e Berg, A. C. (2016). Ssd: Single shot multibox detector. Em *European conference on computer vision*, páginas 21–37. Springer.

Mathias, M., Benenson, R., Pedersoli, M. e Van Gool, L. (2014). Face detection without bells and whistles. Em *European conference on computer vision*, páginas 720–735. Springer.

qqwweee (2018). A keras implementation of yolov3 (tensorflow backend). `https://github.com/qqwweee/keras-yolo3`.

Redmon, J., Divvala, S., Girshick, R. e Farhadi, A. (2016). You only look once: Unified, real-time object detection. Em *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 779–788.

Redmon, J. e Farhadi, A. (2017). Yolo9000: better, faster, stronger. *arXiv preprint*.

Redmon, J. e Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R. e Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Em *Advances in neural information processing systems*, páginas 91–99.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Shrivastava, A., Sukthankar, R., Malik, J. e Gupta, A. (2016). Beyond skip connections: Top-down modulation for object detection. *arXiv preprint arXiv:1612.06851*.

Stauffer, C. e Grimson, W. E. L. (2000). Learning patterns of activity using real-time tracking. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):747–757.

Tian, Y., Luo, P., Wang, X. e Tang, X. (2015a). Deep learning strong parts for pedestrian detection. Em *Proceedings of the IEEE international conference on computer vision*, páginas 1904–1912.

Tian, Y., Luo, P., Wang, X. e Tang, X. (2015b). Pedestrian detection aided by deep learning semantic tasks. Em *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, páginas 5079–5087.

Viola, P. e Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Em *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, páginas I–I. IEEE.

Vu, T.-H., Osokin, A. e Laptev, I. (2015). Context-aware cnns for person head detection. Em *Proceedings of the IEEE International Conference on Computer Vision*, páginas 2893–2901.

Zeng, C. e Ma, H. (2010a). Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting. Em *Pattern Recognition (ICPR), 2010 20th International Conference on*, páginas 2069–2072. IEEE.

Zeng, C. e Ma, H. (2010b). Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting. Em *Pattern Recognition (ICPR), 2010 20th International Conference on*, páginas 2069–2072. IEEE.

Zhang, L., Lin, L., Liang, X. e He, K. (2016). Is faster r-cnn doing well for pedestrian detection? Em *European Conference on Computer Vision*, páginas 443–457. Springer.

Zhang, S., Bauckhage, C. e Cremers, A. B. (2014). Informed haar-like features improve pedestrian detection. Em *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 947–954.