The use of deep learning for person detection and gender classification using RGB images to support the visually impaired

H.A. Jekel*

University of Twente, Faculty of Engineering Technology and Applied Mechanics, Drienerlolaan 5, 7522 NB, Enschede, The Netherlands
**h.a.jekel@student.utwente.nl*

ABSTRACT: This paper presents the deep learning approach to help the visually impaired in an object detection task: recognising the gender of people in their proximate surrounding. By use of images coming from a RPI WWCAM2 monocular camera, the person is first detected , i.e. localized in the image, and then classified to one of the two genders. In order to perform gender detection in real-time, the use of transfer learning together with a single-stage object detection algorithm was investigated. Based on the number of processed frames per second (FPS) and the mean average precision (mAP), it was concluded that fine-tuning a pre-trained YOLOv4 algorithm on customized versions of the Pascal VOC 2007 dataset and the CelebA dataset is best suited for this task.

Key words: Artificial Intelligence, Deep Learning, YOLO, Object Localization, Gender Classification, Object Detection, Transfer Learning, Visual Impairment

1 INTRODUCTION

In 2017, nearly 253 million people in the world suffered from some form of visual impairment [8]. With white canes and guide dogs currently preventing a large number of head-level and fall accidents [6], their assistance does not reach the level of social scenes and human interactions [7]. Besides directly asking about peoples' gender, the visually impaired rely largely on the frequency of someone's voice and other sound cues to determine their gender. Voice frequencies and sound cues can be misleading, causing awkward social situations for the visually impaired. Moreover, environments such as public transport and public libraries do not always allow for the gathering of this auditory information. Localisation of people and classification based on their gender in an unknown environment would help to fill this social gap.

In this paper, the Deep Learning (DL) approach is investigated to assist the visually impaired - later referred to as 'users' - in their social interactions, while gaining a better understanding of their environment. Using DL, an artificial visual system can be produced that replaces the relevant parts and functionalities of the visual organ and central nervous system. Such systems are currently being explored in the car manufacturing industry in the form of autonomous cars [2]. Applied to the visually impaired, a head mounted camera replaces the sensory organs (eyes) and a single-board computer (SBC) equipped with a trained object detection model capable of detecting people and classifying them, replaces the image processing part of the central nervous system. The main challenge of such artificial systems is to make them run in real-time.
Besides the need for real-time running speed, the accuracy is also of high priority. With these two metrics in mind, the object detection architecture is optimized. It should be noted that gender dysphoria can not be considered by the gender classification system discussed in this paper. The classification is therefore purely based on the biological sex. The main aim of this paper is to find the best DL object detection architecture to accomplish this real-time, accurate processing of Red-Green-Blue (RGB) images received from the person mounted camera.

1.1 Previous solutions to challenges in pedestrian- and face detection

The problems of using DL for gender detection are rather specific and resources from previous solutions are therefore scarce. However, one such resource can be found in face detection, one of the oldest computer vision applications. The computer vision problem in gender detection is very similar to the problem in face- and pedestrian detection, since facial and bodily features are involved in both person localisation and gender classification [19]. Previous face detection algorithms, such as the Viola-Jones detector [12], have greatly stimulated the progression of DL applications in today's object detection systems [19].

The histogram of oriented gradients (HOG) detector became widespread in 2005 as one of the first pedestrian detectors[14] and the integral channel features (ICF) detector was introduced in 2009 as a pedestrian detector [15]. They formed a solid foundation for general object detection in terms of the feature representation and the design of classifiers [19]. The faster, region- based convolutional neural network (RCNN) developed in 2015 was proposed as a robust two-stage pedestrian detector [18]. The Faster RCNN did allow for high detection accuracy, but failed to run in real-time. Recently, one-stage object detectors such as single shot multi-box detector (SSD) [38], RetinaNet [37] and you only look once (YOLO) version 4 [21] are surpassing the two-stage object detectors in terms of accuracy and run in real-time [19].

1.2 Traditional detection methods influences in DL detection methods

Traditional detection methods were built based on handcrafted features [19]. The Viola-Jones (VJ) algorithm achieved real-time human face detection for the first time without constraints using integral images, allowing for window size independent computational cost, feature selection and detection cascades [12] to successively pass strong classifiers that each focus on specific features to eliminate 'negative' input faster. Following the VJ algorithm, the HOG [14] and the subsequent deformable parts model (DPM) [13] brought new highly influential methods to the field of object detection, with DPM representing the peak of traditional detectors. Surpassing these highly influential traditional methods in terms of accuracy, is the DL-based method. There are, however, some concepts from traditional methods that are still present in this modern DL method:

- **Hard negative mining**
  The process of dealing with the high foreground-background class imbalance by only using poorly localized positives as negative examples leading to substantially better results. An instance is considered hard if it surpasses a loss threshold [29].

- **Bounding box regression**
  DPM [13] was the first model to ever use bounding box regression. It uses bounding box regression to fine-tune the bounding box prediction presented by the DPM. Therefore it can be seen as a post-processing step.

- **Non-max suppression**
  A post-processing step to remove the replicated bounding boxes and obtain the final detection result. The most common method to apply non-max suppression is Greedy selection [19]. For a set of overlapped detection's, it selects the bounding box with the maximum detection score, while its neighboring boxes are removed according to a predefined intersection over union (IoU), which is the ratio of the intersection of two images over their union. This process is then iterated for all such sets in a greedy manner. This technique was used in HOG [14] and DPM [13].

2 A DEEP LEARNING BASED GENDER DETECTION SYSTEM

Globally, over a quarter of a billion people suffer from visual impairment [8]. These people struggle to initiate social interactions, as they are often unaware of the location and gender of others. Social interaction is a critically important contributor to good health and longevity [10]. Assistance in localizing and classifying people based on their gender is therefore essential, as current aids such as canes and guide dogs fail to do so [7].

With the help of an artificial visual system consisting of a head mounted camera that sends RGB images to a SBC that is equipped with a model capable of detecting and classifying people, one could detect gender or even identify the person on an image. Such a concept is known as a gender detection - or more generally - as an object detection problem.

In an object detection problem, one strives to find and classify a variable number of objects on an image, see figure 1. The variation originates from the changing number of objects in an image (dynamic environment). This describes the main difference between the object detection and the classification problem. In addition to classification, the main challenge in an object detection problem lies in the object localization task. The object detection problem faces several general challenges and issues, the gender detection problem faces some more specific challenges. A more general challenge, the 'objects under viewpoint variation' challenge, can cause the object to dif-

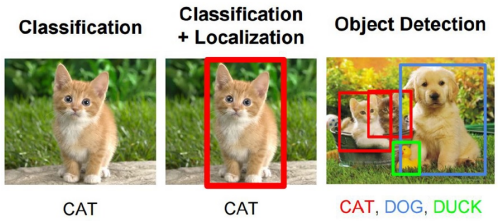

Fig. 1: Object detection: variable number of objects [53].

fer in appearance from various angles. Multi-scale detection is another general challenge, caused by the size difference of objects in an image when they are close by or far away. 'Illuminations' is a third example, where different lighting conditions can cause pixel values to change non-uniformly. 'Object rotation' and 'accurate object localization' are two other examples of general challenges [19]. Specific challenges faced in gender detection are very similar to the specific challenges faced in pedestrian- and face detection, since features of both the face and body are involved in person localisation and gender classification [11]. These specific gender detection challenges include intra-class variation, occlusions, hard negatives and real-time detection. Intra-class variation refers to the uniqueness of human expressions, skin color, poses and movements. Due to the dynamic behaviour of people, face occlusions are common, making it difficult to detect the face. Hard negatives are encountered when backgrounds contain pedestrian-like features [11]. Lastly, real-time detection is essential, as in dynamic environments people move around changing positions constantly.

The DL approach is used in this paper to solve both the localization and classification challenges. A DL object detection model is a complex mathematical function that maps RGB images to bounding boxes and class predictions. This complex function is learned through a process called training, in which the detectors predictions are compared to the ground truth label. The comparison leads to the computation of a cost when the prediction and ground truth are not identical. The optimization problem in which the detectors' weights are adjusted as to minimize the cost then leads to a function that is able to perform the aforementioned mapping. To train the model, a set of annotated RGB images collected by the RPI WW-CAM2 monocular camera could be utilized to train the DL object detection model. However, considering the high costs in terms of time and effort to create such a dataset, a novel methodology presented in Section 5.2 is considered.

3 METHODOLOGY

This section starts off with an analysis of both two-stage and one-stage object detectors as a potential candidate for the DL gender detection model. It then substantiates the convergence to the cutting edge one-stage detectors, for which the numerical results are given in section 4.

3.1 Multi-stage detectors

One possible object detection architecture for a gender detection system is the RCNN architecture (2014) proposed by [18], figure 3. RCNN uses the divide and conquer approach, as it divides the complex object detection task in 3 stages, see figure . First, a selective search algorithm is used to propose regions of interest (RoI) in the image that might contain an object by adding a bounding box. The region proposals are then warped (re-scaled) and fed into a convolutional neural network (CNN) to extract their high level features (backbone). These features are further processed in a few fully connected layers (neck). Finally, the regions are classified using a support vector machine (SVM) classifier, also referred to as large margin classifier. To refine the initially proposed bounding box coordinates of the object, bounding box regression is applied. The classification part, together with the bounding box regression part of the architecture, is referred to as the head of the architecture. This two-stage sparse architecture is displayed in figure 2. With a post-processing step called non-max suppression, the redundant bounding boxes that detect the same object are removed [18, 38, 37, 25, 26, 27, 28].

A big limitation of the RCNN is the inability to process images of different sizes. RoIs need to be re-scaled before the RCNN can process them, as the fully connected layers in the neck require a fixed input size [40]. SPP net uses pooling layers with a fixed amount of bins (independent of the input size) to process the CNN's output feature maps, concatenating them, resulting in fixed size output, see figure 4. This is faster than cropping and warping like RCNN, because the feature maps only need to be calculated once as the RoIs are mapped to the feature

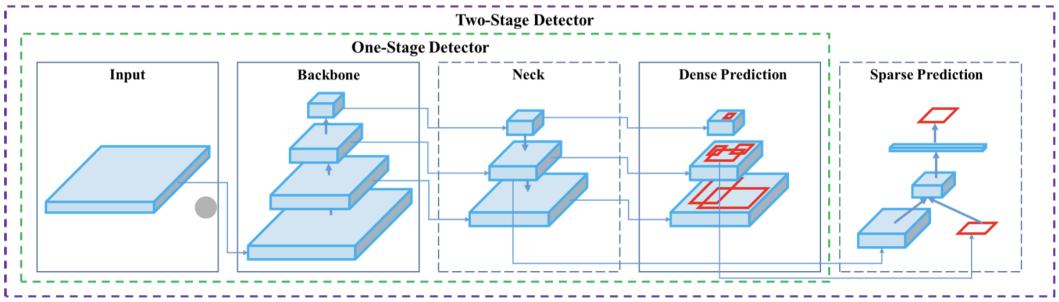

Fig. 2: Architecture of two-stage and one-stage object detectors [28].

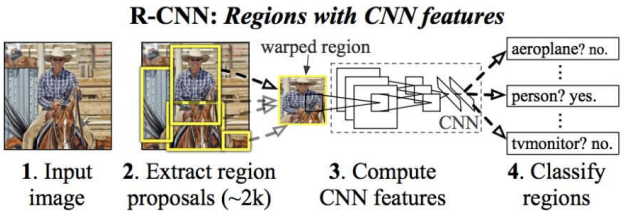

Fig. 3: RCNN: devide and conquer [16].

maps using the sub-sampling ratio (spatial scale ratio from input image to feature map). However, the SPP layer simultaneously presents an additional challenge, as back-propagation (a process needed to train the detectors weights) through the SPP layer is limited due to the multiple filters used. Therefore, when implementing the SPP layer, one requires multi-stage training: training the backbone separately from the neck and the head [40].

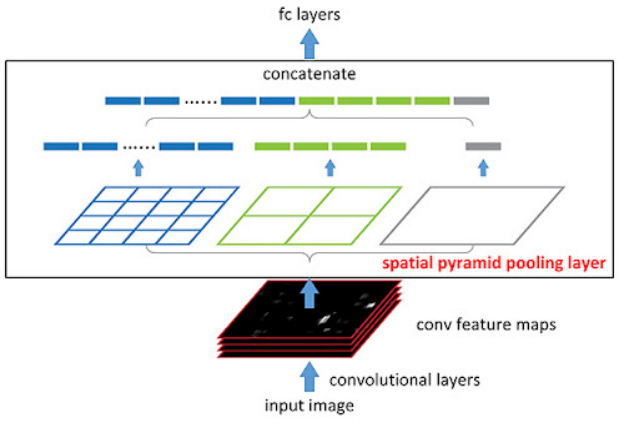

Fig. 4: RCNN: devide and conquer [40].

A more promising base architecture for gender detection is the Fast RCNN (2015), as it solves the back-propagation limitation faced by SPPNet by removing all spatial pooling filters, except one.[17] This allows for simultaneous training of the backbone, neck and head, while only calculating feature maps once. Fast RCNN also introduces multi-task loss, including both localization and classification. In the multi-task loss function, cross entropy loss and smoothL1 loss are combined, resulting in a single optimisation problem.

Another possible base architecture for a gender detection system is the Faster RCNN (2015) [18]. The autors introduced the region proposal network (RPN) as a substitution for the selective search algorithm [18]. The structure of the RPN is supported by a convolutional implementation of sliding windows [48], allowing for convolutionally implemented, fully connected layers [24]. Faster RCNN brought another innovation to the DL detection field, namely proposing an anchor of fixed scales and aspect ratios based on the size and shapes of objects contained in the dataset. The region proposal network proposes these anchor boxes of different scales and aspect ratios for each of the sliding windows and additionally predicts if it contains an object. If it does, it becomes a RoI and is fed into the second stage where classification and bounding box regression is applied to the anchor boxes containing objects, refining their localisation (sparse prediction, figure 2). Proposing multiple anchor boxes of different scale and aspect ratios allows the output units of the RPN to specialize for a certain shape and allows for the detection of multiple objects in one window.

In conclusion, some of the redundant steps of the multi-stage objet detectors were solved, e.g. by replacing warping by spatial pyramid pooling [40], allowing for shared computation in the second stage and by replacing the selective search algorithm by the RPN in the first stage. For real-time applications, however, the multi-stage detector architecture remains

insufficient [25].

3.2 *One-stage detector*

The one-stage detector combines both localization and classification in a single CNN. The general idea of a one-stage detector is to first feed an image into a CNN (the backbone) and combine its extracted features maps in the neck. Lastly, the head assigns bounding boxes to the objects and allocates class probabilities, predicting which class is most likely contained within the bounding box. This dense one-stage object detection architecture is displayed in figure 2.

Another possible base architecture for gender detection was proposed by the autors of the YOLO algorithm (2016), see figure 5. They recognized that the speed limitations of Faster RCNN can be traced down to its two-stage architecture. They were the first to explore the idea of merging the two-stage object detection architecture into a single stage, only looking once at the input image [25]. The YOLO architecture is very similar to the architecture of the first stage of the Faster RCNN two-stage detector. The big difference from Faster RCNN however, is that YOLO adds class predictions directly which removes the need for a second stage. Another difference is that YOLO does not implement the multiple anchor boxes for each window, as introduced by Faster RCNN. This causes the YOLO algorithm to struggle with detecting small objects. The authors decided to tackle the problem of multiple-scale detection in their second version, YOLOv2 (2017), by using multiple anchor boxes of different size and scale for each window [26], as previously proposed by Faster RCNN. Although able to process images in real-time, the YOLOv2 architecture could not match the Faster RCNN accuracy.

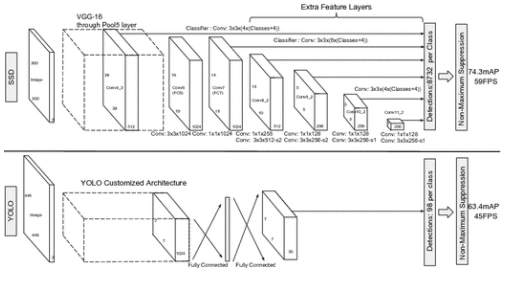

Fig. 5: The SSD architecture (top) and the YOLOv1 architecture (bottom) [38].

3.3 *SSD*

The base architecture of the SSD, figure 5, can be considered as a serious candidate of the gender detection system, as it represents a state of the art object detection architecture. Similar to YOLOv2, SSD utilizes the Faster RCNN-proposed bounding boxes of different scales and aspect ratios to optimize the accuracy. Additionally, SSD proposes the idea of basing the prediction on intermediate feature maps together with the output feature map. YOLOv1 and YOLOv2 can only predict based on the output feature map. This is denoted in figure 5 with the line connections between the intermediate feature maps and the output layer for SSD and the line connection between the output feature map and the output layer for YOLOv1 and YOLOv2. To solve the class imbalance problem as stated in section 1.2, SSD picks the negative predictions with the highest loss and makes sure the ratio between the picked negative and positive predictions is at most 3 to 1. Hence, only the gradients of a very small part of samples (those with the largest loss values) will be back-propagated. This is process is called online hard example mining (OHEM). [29]

3.4 *RetinaNet*

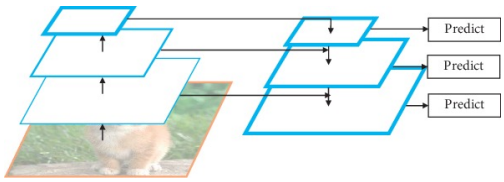

Fig. 6: RetinaNet neck: Feature pyramid network [44].

Another serious candidate is RetinaNet. RetinaNet was the first one-stage object detector to surpass the two-stage object detectors in accuracy. RetinaNet uses a feature pyramid network (FPN) neck on top of a ResNet backbone (see figure 6) [37, 44]. The

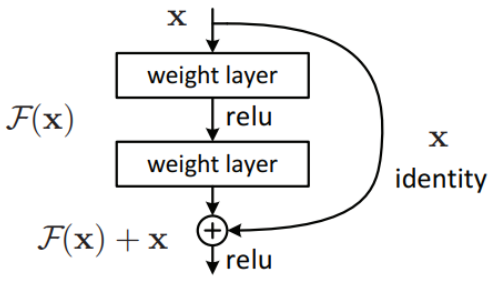

Fig. 7: Residual Learning: a building block [39].

ResNet architecture added a revolutionary component

to CNN's called the skip-connection, (see figure 7) [39]. It feeds forward information, solving the vanishing/exploding gradient problem faced in deep neural networks. Using skip connections the network can easily learn the identity function, allowing deeper CNN's to continuously improve performance. The FPN neck uses transpose convolutions to build a spatially larger feature map from the spatially smaller deeper feature map. It then combines the enlarged feature map with feature maps from earlier layers of the same size and making predictions based on their combination [44]. The component of RetinaNet that

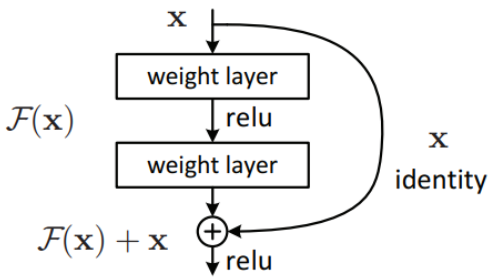

Fig. 8: Focal loss [37].

made it a breakthrough detector was the focal loss, see figure 8. Focal loss addresses the extreme imbalance between foreground and background classes during training. The autors reshape the standard cross entropy loss so that it will put more focus on hard, miss-classified examples.

3.5 YOLOv4

YOLOv4 claims the last position as a serious candidate for the gender detection system. It consists of a CSPDarknet53 backbone, adding a SPP module (figure 4) and replacing the FPN neck by a PANet [41] (figure 11) neck, while retaining the YOLOv3 head. Cross-stage partial connection is a DenseNet [32] based architecture. Densnet, figure 9, uses ResNets' idea of skip connections and connects everything together to facilitate backpropagation during training. [32] CSP uses the same filosofy, but removes the duplicate gradient information generated by DenseNet. [36]. The PAN neck is an improved version of the FPN neck, as can be seen from figure 6 and figure 11 [41].

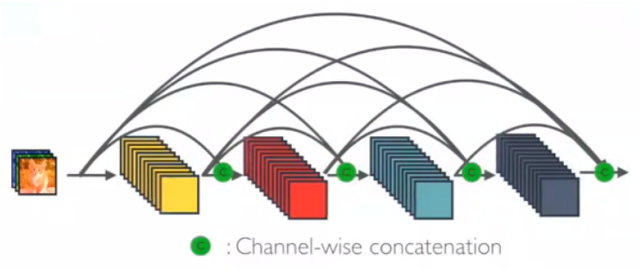

Fig. 9: DenseNet architecture [32].

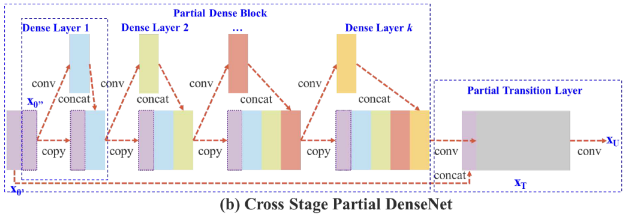

Fig. 10: CSP [36].

4 NUMERICAL RESULTS

In this section, a numerical analysis of the three state of the art one-stage object detectors - SSD, RetinaNet and YOLOv4 - is outlined based on a literature search. This will result in the selection of the most appropriate one-stage detector for implementation in gender detection.

As stated in the introduction, the gender detection system should be optimized for both speed and accuracy. To measure the accuracy, it is important to first think about what might occur in different detection failure scenarios regarding gender detection. When a false positive occurs, the user might start a conversation using the wrong gender, or even worse, the user might start talking to a brick wall. To prevent this, the precision of the detector should be optimized (figure 12). On the other hand, it is important to think about what would happen if the detection system fails to detect someone. With this failure, an opportunity for a social interaction is lost, failing to pursue the main goal: facilitating users in social interactions. To make sure that the system is useful, recall should be optimized (figure 12). Another object detection measure for accuracy is the mean Average Precision (mAP). When optimizing the mAP, one optimizes both precision and recall. The mAP is therefore well suited as a measure of accuracy for gender detection. It is determined by computing the average of the areas below the precision and recall curve, for all ob-

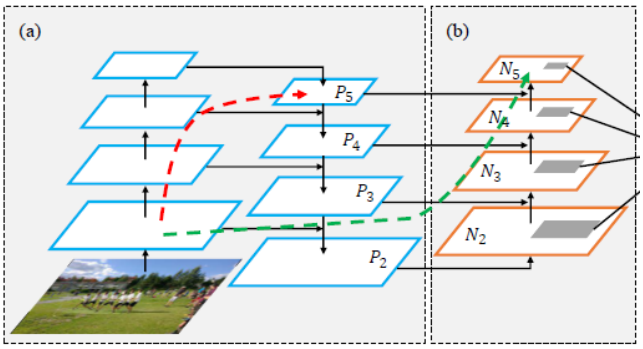

Fig. 11: PAN [41].

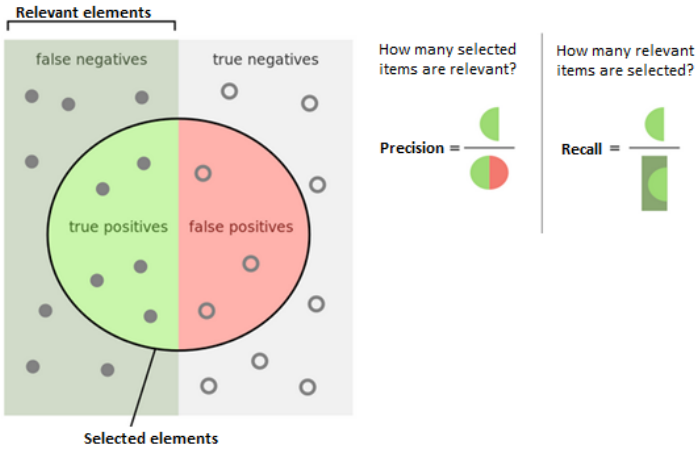

Fig. 12: Precision and recall [23].

ject classes, to a certain IoU true positive threshold. Specifically, here the mAP@0.5:0.05:0.95 is used. To compute mAP@0.5:0.05:0.95, one takes the average of the mAP's for the IoU thresholds ranging from 0.5 to 0.95 with step size 0.05. The FPS is a common metric to compare the processing speeds of different object detectors. In order to compare the three state of the art object detectors, their mAP@0.5:0.05:0.95 (in table 1 referred to as AP) and FPS are displayed in table 1. Here, the detectors are trained and tested on the MS COCO dataset containing 350.000 labeled objects and 80 different object categories. Table 1 shows

Table 1: Comparison of the speed and accuracy of SSD, RetinaNet and YOLOv4 on the MS COCO dataset (with batch=1 without using tensorRT).

Method	**Backbone**	**Size**	**FPS**	**AP**
SSD	VGG-16	512	22.0	28.8
RetinaNet	ResNet-101	500	11.1	34.4
YOLOv4	CSPDarknet-53	512	31.0	43.0

that YOLOv4 is the best object detection architecture, as it performs better in both FPS and AP compared to the other object detectors.

5 METHODOLOGY FOR FURTHER RESEARCH

Since this papers' results are based exclusively on literature research, this section discusses how to perform further research to select the best of the three state of the art one-stage object detector architectures for gender detection.

5.1 Experiment

To quantitatively compare the three single-stage architectures for gender detection, they must all be prepared identically. First, weights of each single-stage detector that are pre-trained on the MS COCO dataset [49] will be collected from the authors. As the first layers of the backbone always learn to extract general, low level features (e.g. edges, circles) almost independent of the dataset, the pre-trained weights allow for the model to use these low-level features to converge to the gender dataset faster. During fine-tuning, the subsequent layers will learn to extract high-level features which are able to distinguish between male and female. Before fine-tuning, the head of the pre-trained model is removed and a gender detection head, containing only the male and female object classes, is added. The detector is then fine-tuned on a customized gender detection dataset, as is discussed in section 5.2. The head substitution and the fine-tuning process together is referred to as transfer learning [31]. Freezing certain layers before fine-tuning is another common practice which could be exploited. To determine a suitable number of epochs for fine-tuning, the pre-trained YOLOv4 network is fine-tuned first on 30 epochs of the training set as fine-tuning requires fairly little training. A plot of the loss during training then allows to determine an appropriate number of epochs to train the models. During training, all 3 loss functions are plotted, allowing to check for proper training behaviour. To quantitatively compare the 3 object detectors both FPS and mAP are used. Following training, a testing set is used to compare the models on their mAP and FPS.

5.2 Dataset

To make sure that the final gender detection system best assists the users in their social interactions,

it is important to apply both a model-centric and a data-centric approach to the design of the system. The results section of this paper compares different models on their speed and accuracy given a fixed dataset: a model-centric approach. However, it is essential to also have a data-centric approach, emphasizing the creation of an efficient learning dataset. This translates to the current application by using the application-specific camera to record social situations that users will encounter. Besides the privacy issues regarding the personal data of users, one needs to consider the high costs in terms of time and effort to create such a dataset. For each image, an annotator needs to draw a bounding box around the identified people and classify them as male, female or unknown. Time constraints and limited resources may not allow for creation of such a dataset. The only available application-specific datasets seem to be those created by de Araujo and colleagues [22]. The contribution of said authors [22] exists of two new gender datasets. First they customized the Pascal VOC 2007 [50] dataset. VOC originally contains 20 object classes, including people captured at different poses and occlusions. The authors first removed all classes except for the person class and then added labels 'man', 'woman' and 'undefined' to these bounding boxes. This dataset consists of 4192 annotated images with 4381 instances of men, 3210 instances of women and 3083 instances of persons with undefined gender. The dataset comes with challenges, as images usually contain a high variation of poses, occlusions, intersections and multiple instances. As this dataset is small in comparison to regular fine-tuning datasets [23], the authors of [22] also provided a customized version of the CelebA [52] dataset. This dataset contains faces with their annotated gender and bounding boxes that surround the faces. However, as bodily features might also be useful to detecting gender, the authors ran a person detector over the CelebA dataset to adjust the bounding boxes. For training and testing, they analyzed the use of different proportions of each dataset, concluding that a 50-50 distribution is optimal. A K-folds validation scheme, where a random selection from the customized CelebA and VOC datasets is used for each epoch, would be most appropriate in training. As the detector learns using labeled data, the corresponding estimation problem can be described as supervising [23]. As discussed, the best option for training an algorithm is currently described by [22] and it seems clear that improvements can be made on the data side of things. Therefore, it is proposed that further research should contribute newly annotated datasets, taking, for example, the persons in the MS COCO dataset and annotating their gender.

5.3 Hardware implementation: the artificial visual system

Here, a set-up of the artificial visual system prototype is discussed. The best gender detection model is uploaded onto a Raspberry Pi (RP) 3. A Raspberry Pi 3 is chosen to be the SBC as it is a small, portable and concealable device. A head mounted RPI WWCAM2 monocular camera feeds the RGB images through a RP camera serial interface to the model. The gender detection model then sends the detection information to an audio program that will forward key words to the user via a wired earbud. To power the system, a portable power module is linked with the Raspberry Pi. The full gender detection models are large in size and computational cost. If the models appear to be too large after initial experimentation, the models will be post-processed using TensorflowLite [51] to reduce their size and computational cost.

6 CONCLUSION

The main aim of this thesis was to find the best deep learning object detection architecture for a gender detection model to aid the visually impaired in their social life. A literature comparison of the three state of the art object detection architectures showed that, from the three one-stage object detection architectures, the YOLOv4 architecture outperforms all other architectures. Therefore, based on this research, the YOLOv4 architecture seemingly has the most potential for the artificial visual system. To support this conclusion, further research should aim to modify the head of the three state of the art object detection architectures to the gender detection task, train and test the modified architectures on a representative gender dataset, and finally compare the numerical results.

7 ACKNOWLEDGEMENTS

I would like to express my gratitude to Prof. Dr. Ing. B. Rosic, for encouraging me to develop my research skills individually, inspiring me to embrace the field of robotics with a mathematical view and for providing me with large amounts of feedback on my aca-

demic writing in the last stages. I would also like to thank V. Rokx-Nellemann and T.I.J Jekel for proof reading my paper.

REFERENCES

1. A. Tang, R. Tam, A. Cadrin-Chênevert, W. Guest, J. Chong, J. Barfett, L. Chepelev, R. Cairns, J. Mitchell, M. Cicero, M. Gaudreau Poudrette, J. Jaremko, C. Md, B. Gallix, B. Gray, R. Geis, T. O'Connell, P. Babyn, D. Koff, W. Shabana, Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. In: *Canadian Association of Radiologists Journal, Vol 69*, 120-135 (2018).

2. S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, A survey of Deep Learning Techniques for Autonomous Driving. In: *Journal of Field Robotics, Vol. 37*, 362-386 (2019).

3. W. Serrano, Neural Networks in Big Data and Web Search. In: *Data, Vol. 4, Issue 7*, 1-41 (2018).

4. P. Covington, J. Adams, E. Sargin, Deep Neural Networks for YouTube Recommendations. In: *RecSys'16: Proceedings of the 10th ACM Conference on Recommender Systems*, 191-198 (2016).

5. D. Chatterjee, The Rise of Deep Learning in Radiology: An Overview of Recent Research. In: *International Journal for Research in Applied Science and Engineering Technology, Vol. 7*, 2353-2361 (2019).

6. R. Manduchi, S. Kurniawan, Mobility-related accidents experienced by people with visual impairment. In: *Insight: Research and Practice in Visual Impairment and Blindness, Vol. 4*, 44-54 (2011).

7. P. Craigon, P. Hobsson-West, G. England, C. Whelan, E. Lethbridge, L. Asher, "She's a dog at the end of the day": Guide dog owners' perspectives on the behaviour of their guide dog. In: *Plos One, Vol. 12*, (2017).

8. R. Bourne, S. Flaxman, T. Braithwaite, M. Cicinelli, A. Das, Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. In: *The Lancet Global Health, Vol 5, Issue 9*, e888-e897 (2017).

9. S. Kothiya, K. Mistree, A review on real time object tracking in video sequences. In: *Electrical, Electronics, Signals, Communication and Optimization*, 1-4 (2015).

10. D. Umberson, J. Karas Montez, Social relationships and health: A flashpoint for health policy. In: *Journal of health and social behavior, Vol. 51, No. 1*, 54-66 (2010)

11. N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection. In: *computer society conference on computer vision and pattern recognition, Vol. 1*, 886-893 (2005).

12. P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. Vol. 1*, 1-9 (2001).

13. P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models. In: *IEEE transactions on pattern analysis and machine intelligence, Vol. 32, No. 9*, 1627-1645 (2009).

14. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection. In: *computer society conference on computer vision and pattern recognition, Vol. 1* 886-893 (2005).

15. P. Dollár, Z. Tu, P. Perona, S. Belongie, Integral Channel Features. In: *British Machine Vision Conference* (2009).

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

17. R. Girshick, Fast r-cnn. In: *Proceedings of the IEEE international conference on computer vision*, 1440-1448 (2015)

18. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39*, 1-14 (2015).

19. Z. Zou, Z. Shi, Y. Guo, J. Ye, Object detection in 20 years: A survey. In: *arXiv:1905.05055*, (2019).

20. X. Yang, Y. Wang, R. Laganiere, A Scale-Aware YOLO Model for Pedestrian Detection. In: *International Symposium on Visual Computing*, 15-26 (2020).

21. K. Zheng, M. Wei, S. Li, D. Yang, X. Liu, Pedestrian Detection in Driver Assistance Using SSD and PS-GAN. In: *Journal of Autonomous Intelligence, Vol. 2, Issue 3*, 9-18 (2019).

22. Z. de Araujo, F. Luis, C. Jung, Real-time gender detection in the wild using deep neural networks. In: *SIBGRAPI Conference on Graphics, Patterns and Images*, 118-125 (2018).

23. A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, California (2019).

24. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431-3440 (2015)

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

26. J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263-7271 (2017)

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

28. A. Bochkovskiy, C. Wang, H. Liao, Yolov4: Optimal speed and accuracy of object detection. In: *arXiv:2004.10934* (2020)

29. A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* 761-769 (2016).

30. C. Wang, I. Yeh, H. Liao, You Only Learn One Representation: Unified Network for Multiple Tasks. In:

arXiv:2105.04206 (2021).

31. D. Wu, S. Lv, M. Jiang, H. Song, Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. In: *Computers and Electronics in Agriculture, Vol. 178*, 105742 (2020).

32. G. Huang, Z. Liu, L. Van Der Maaten, K. Weinberger, Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700-4708 (2017).

33. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132-7141 (2018).

34. Z. Yao, Y. Cao, S. Zheng, G. Huang, S. Lin, Cross-iteration batch normalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12331-12340 (2021).

35. S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International conference on machine learning*, 448-456 (2015).

36. C. Wang, H. Liao, Y. Wu, P. Chen, J. Hsieh, I. Yeh, CSPNet: A new backbone that can enhance learning capability of CNN. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 390-391 (2020).

37. T. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*, 2980-2988 (2017).

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

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

40. K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition. In: *IEEE transactions on pattern analysis and machine intelligence, Vol. 37, No. 9*, 1904-1916 (2015).

41. S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8759-8768 (2018).

42. S. Woo, J. Park, J. Lee, I. Kweon, Cbam: Convolutional block attention module. In: *Proceedings of the European conference on computer vision*, 3-19 (2018).

43. D. Misra, Mish: A self regularized non-monotonic neural activation function. In:*arXiv:1908.08681* (2019).

44. S. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117-2125 (2017).

45. A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications. In: *arXiv:1704.04861* (2017).

46. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510-4520 (2018).

47. M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*, 6105-6114 (2019).

48. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks. In: *arXiv:1312.6229* (2013).

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

50. M. Everingham, L. Van Gool, C. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 Results. In: *http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html* (2007).

51. P. Warden, D. Situnayake, Tinyml: Machine learning with tensorflow lite on arduino and ultra-low-power microcontrollers. Publisher: *O'Reilly Media* (2019).

52. Z. Liu, P. Luo, X. Wang, X. Tang, Deep Learning Face Attributes in the Wild. In: *Proceedings of International Conference on Computer Vision* (2015)

53. A. Khan, S. Al-Habsi, Machine learning in computer vision. In: *Procedia Computer Science, Vol. 167*, 1444-1451 (2020)