Improving Object Detection in Real-World Traffic Scenes

Waqar Khan\textsuperscript{1}, Fang Liu\textsuperscript{1}, and Marta Vos\textsuperscript{2}

\textsuperscript{1} Wellington Institute of Technology Ltd., Wellington, New Zealand
wkha011@aucklanduni.ac.nz

\textsuperscript{2} Whitireia Community Polytechnic Ltd., Wellington, New Zealand

Abstract. Single Shot Multi-Box Detector (SSD) is a well-known object detection algorithm. It can detect 20 different types of objects making it suitable for an object detector for traffic scenes. In a real-world traffic scene, objects can appear in different sizes and pose different details. This can potentially lead to false detections made by an SSD. Depending on how input information (image) is provided to SSD (leading to a proposed SSD model), the accuracy of the proposed model can vary. The overall objective of this study is to evaluate different SSD models while examining accuracy of object detection where the object type is only a vehicle.

This study is derived from human vision. Where, an object is easily identifiable in a sharper image with brightness than a blurry one with darkness. Based on these assumptions hypotheses were created, based on which SSD based models were proposed. Comparison based on true positives and false positives was performed and the winner was identified by using the Enpeda. Image Sequence Analysis Test Site (EISATS) stereo image barriers dataset set 9.

Keywords: Object detection \cdot Ground truth \cdot True/false positive \cdot Precision \cdot Recall \cdot Win count

1 Introduction

According to the World Health Organization (WHO), the road accident casualties were more than 1.35 million in year 2016 [1]. To avoid such scenarios, computer vision based safety systems can assist in identifying hazards and either apply brakes or warn the driver about imminent collision, so that he/she could apply brakes [2]. These safety systems have to track an object of interest with time to estimate their trajectory, and after they are confident, can issue the warning to the driver. Apart from the stereo measurement inaccuracies [3], there are also object tracking inaccuracies, particularly in scenarios where objects keep on crossing each other. And, drift could occur.

In such a challenging scenario, a hybrid approach is often adopted, where object is detected in every frame, and each detection is compared with previously tracked object, to identify if it is the same object or a different one. This process is also known as data association [4]. Data association is particularly important when the visual tracker fails.
due to different factors like occlusion, etc. So, in other words, object detection is not only important to detect an object in the first frame, however, continuous detections are also essential.

In other words, this implies that when the vehicles in the traffic scene are always moving (like in EISATS set 9 [5]). Then, the accuracy of detector is determined by its performance in every frame/image. A moving camera in particular depicts a different scene for the same object. It is because the pose of the object of interest relative to the camera capturing the scene is always changing with time. It is true, provided that both the object of interest and the ego-vehicle (with camera mounted on it) are moving at a different velocity in world co-ordinates.

In the context of object detection, only left camera images are sufficient from EISATS set 9 which consists of 400 greyscale images. An object detector identifies the type of object and its location. Therefore, the performance of a detector is based on how accurately it classifies an object’s type as well as its location in the 2D image. To evaluate the performance, data with ground truth (GT) is used. The evaluation output could be true positive (TP), where detected object (type and location) matches with the actual (type and location) of object as indicated by GT. It could be false positive (FP), where the detection type or location do not match with GT. It could be false negative (FN), where there is no detection, however, an object exists in GT. And, finally, it could be true negative (TN), where no object is detected and no object exists at that location in GT either.

FNs often occur when objects are smaller in size as they are farther from the camera, or when the objects of interest are occluded. Because EISATS set 9 dataset consists of a sequence of images captured over time, this allows for TPs, FPs and FNs to be critically evaluated.

2 Literature Review

Before the introduction of convolutional neural networks (CNN), object detection relied on basic features as they were supported by limited processing powered devices. For example, Viola and Jones developed a detector called VJ detector which could detect faces in real-time [7]. They used sliding window operations to compute features in all locations and at different scales.

In 2005, N. Dalal and B. Triggs built a pedestrian detector based on Histogram of Oriented Gradients Feature Descriptor (HOG descriptor) [8]. HOG was generalized and was able to detect various objects of different sizes. It was able to achieve that because, it resized the input window to match with the fixed filter window size and used it to compute the feature descriptor. This approach has been adopted by several later algorithms including Deformable Part-based Model (DPM). DPM was first proposed by P. Felzenszwalb et al. in 2008 [9], which was further improved by R. Girshick in 2010 [10]. DPM won series of detection challenges from year 2007 to 2009. In the training phase, DPM breaks down an object into its parts, while in the testing phase it tries to assemble detections to those parts. The development of DPM also involved development of multi-instance learning as well as bounding box regression.

In 2012, a revolutionary deep learning method: CNN was proposed by A. Krizhevsky, I. Sutskever, and G. E. Hinton [11]. It was able to classify more than 1 million images
in an ImageNet training set, consisting of 1000 different classes. Its error rate of 39.7 percent was significantly lesser than others proposed at the same time.

Typically, in a CNN, all pixel values of an input image are processed via multiple hidden layers. These layers consists of several convolutional layers, which allows for the extraction of features like edges or corners etc. These layers also consist of max pooling, which is a process adopted to reduce the dimensions of data, leading to reduction in processing time.

In 2014, Girshick et al. proposed the Regions with CNN features (R-CNN) which is a two-stage detector [12]. Other, two stage detectors followed including Fast R-CNN [13], and Faster R-CNN [14].

In 2016, W. Liu et al. proposed one-stage detector named SSD [15]. SSD needs only a single shot to detect multiple objects in an input image. Due to this, SSD is much faster than two stage detectors. For example, SSD300 achieved an mAP of 74.3 percent, with 59 frames per second (FPS) while SSD500 achieved 76.9 percent mAP at 22 FPS. Both of these outperform Faster R-CNN which had 73.2 percent mAP and 7 FPS. While YOLOv1 had 63.4 percent mAP at 45 FPS.

3 Methodology

Although we chose to use the EISATS set 9 dataset, which is ideal for evaluating the effect of size of any object as well as brightness conditions, it does not come with GT. Our first step was to create a GT dataset for the object type vehicle/car for all 400 frames. For the purpose of evaluation, GT data had to capture the bounding box co-ordinates for each object in each frame. This was a completely manual exercise at least without any object tracking to identify the object location in subsequent frames. Nevertheless Python scripting was used to capture the ground truth into a text file.

Algorithm 1 describes the steps taken for capturing GT for 400 images of EISATS set 9.

Algorithm 1: GT collection from raw images of dataset

```plaintext
Set car_ground-truth_list
Set truck_ground-truth_list
for i in range (1,401):
    Read image i
    WHILE user inputs ‘n’ to log in new object’s box
        IF user indicates to capture a car by pressing ‘c’
            Write to file mouse_coordinates_car
        ELSEIF user indicates to capture a truck by pressing ‘t’
            Write to file mouse_coordinates_truck
        ELSEIF user indicates to go to next image by pressing ‘n’
            BREAK
    ENDIF
Write top-left corner as (x1, y1) and bottom-right corner as (x2, y2).
```

Write top-left corner as (x1, y1) and bottom-right corner as (x2, y2).
The evaluations were based on the performance of SSD based models on the given dataset. Each model was proposed on top of the previous one. So, overall this research was an iterative process, as illustrated in Fig. 1.

Model 1: Raw greyscale images were processed by SSD, the SSD detection accuracy was measured against the GT using the FPs, TPs and FNs for each image of the dataset. Several experiments were conducted to create a hypothesis for proposing an appropriate model for stage 2. The detection accuracy for each image constitutes result 1.

Model 2: SSD processed resized greyscale images using the hypothesis from stage 1. Experiments similar to stage 1 were repeated to create a further hypothesis and an appropriate model was proposed for stage 3. The detection accuracy lead to result 2.

Model 3: SSD used previous hypothesis. This model also used pre-processed images with image sharpening and histogram equalisation respectively.

A new model iSSD (with two variance iSSDv1 and iSSDv2) was proposed based on the iterations above.

Since the detection bounding box location could be anywhere for a given object and it potentially would not occupy the same region as GT bounding box. So, intersection over union (IoU) ratio is computed to identify if the detection (of the same object type) is TP or FP. The IoU process is further illustrated in Fig. 2.

Because GT is labelled manually, without any correlation from one image to the following with time. Therefore, a loose IoU measure of 0.5 is chosen to complement the
tightness of GT production. Usually only when IoU \( \geq 0.5 \), can a detection be considered as TP. If IoU \(< 0.5\), then that detection is considered as a FP. Furthermore, all the detections, where the categories do not belong to GT categories, are considered as FPs as well. For example, if a detection is categorised as a bus or a horse, but there are no such categories in GT, they will be deemed as FPs.

![Fig. 2. IoU presents the level of similarity between SSD detection and GT.](image)

Furthermore, when two or more detections overlapping with each other, and, both have an IoU greater or equal 0.5 with the same object. Then, the one with larger IoU is considered as TP, and other/s is/are deemed as FP for the given object.

### 3.1 Precision and Recall

In computer vision, precision (also known as positive detection value) equals the fraction of TPs from all the retrieved object instances [16]. It is a metric to show precision of detecting instances. The standard formula of precision is

\[
P_{CT_{old}} = \frac{TP}{TP + FP}
\]

Where CT stands for confidence threshold.

On the other hand, recall which is also known as sensitivity, or TP rate, is the fraction of the total amount of correct detections compare to the actual total instances. This also results in computing a percentage. The standard formula of recall:

\[
R_{CT_{old}} = \frac{TP}{TP + FN}
\]

Precision and Recall together are a common pair of metrics used to evaluate the performance of object detection algorithms.

However, a new precision and new recall were proposed for this study.

**New Precision and New Recall**

To avoid 0-division error, when there are no detections in the given image, the formula was slightly modified to

\[
P_{CT_{new}} = \frac{TP}{TP + FP + 0.00001}
\]
Similarly, to avoid undefined recall values due to no TPs in the given image. The formula was slightly modified to

\[ R_{new}^{CT} = \frac{TP + 1}{TP + FN + 1} \]

To be specific, when comparing different models against each other, \( R_{old}^{CT} \) does not appropriately serve in distinguishing between them. This is further described in the following example scenario:

Image A has one car in GT, while image B has ten cars in GT. If TP = 0 for both images, then based on the examples, the performance evaluator should indicate that a detector performed far more poorly on image B with FN = 10 than on image A with FN = 1. However, \( R_{old}^{CT} \) for both image A and B is equal and zero. Therefore, a slightly modified version is proposed from \( R_{old}^{CT} \) to \( R_{new}^{CT} \). Based on this, with \( R_{new}^{CT} \) it gives 0.5 and 0.0909 for images A and B respectively. Such difference becomes critical to cross evaluate several models against each other for each image.

Furthermore, in vision-based driver safety systems, every FN means a hazard is missed by the system, which is costly. More FN are more costly than fewer FN. So, it is better to distinguish this situation using \( R_{new}^{CT} \).

### 3.2 Score System

Due to the complexity of image inputs and result outputs, it is difficult to compare the performance of different models independently. In [17], W. Khan et al. proposed a win count system to compare models. Based on their method, in this research, a slightly different score system was designed, named “scorePR”, for precision and recall. Where, for each image one model may or may not be scored in terms of recall and precision. After the winning score for precision or recall for each image has been collated, a total score is computed to find out which candidate/model gets higher score for the entire dataset in relation to either precision or recall.

For example, three situations described based on three example images and for each three candidates are evaluated, where each candidate is just an SSD model:

- **In image 1**, no candidates win because all candidates are equal due to zero precision. All candidates get zero mark.
- **In image 2**, only candidate 1 wins because only candidate 1 has the highest precision value. Candidate 1 gets one mark; the others get zero mark.
- **In image 3**, both candidate 1 and candidate 2 win, because they both have the equal highest precision value of 1. Candidate 1 and candidate 2 get one mark, while others get zero mark.

After applying the same approach on each image, scores can be summed for each candidate. The total in the example becomes, two, one and zero for candidates 1, 2, and 3 respectively. So, based on this candidate 1 is the winner based on score precision or scoreP. Similarly, score recall is computed and together they are called scorePR.

The same system was applied to recall. Based on this, the overall performance could be evaluated and compared quantitatively, rather than qualitatively.
### 3.3 SSD Based Models

Three models are evaluated. SSD is evaluated against improved SSD which adopts pre-processing to images in the form of histogram equalization. This model is called iSSDv1. Whereas, the iSSDv2 incorporates histogram equalization as well as image sharpening. Figure 3 illustrates histogram equalization on one of the images from EISATS set 9.

The histograms in Fig. 3 describe the grey scale intensity frequencies i.e. from 0 (dark) up to 255 (white). In Fig. 3 (A), it is clear that by default the image has more dark pixels than white. Whereas, after histogram equalization, when the pixel intensities are relatively equally distributed, the dark regions became brighter (see Fig. 3 (B)).

![Image 400 of EISATS set 9 with (right) and without (left) histogram equalisation.](image)

### 4 Results and Discussion

The evaluation is also a three stage process. In the first stage, the goal is to identify the best CT for the given dataset. Once this is identified, in the later experiments the chosen CT is kept as constant, and suitable image resolution is identified. In the final experiment, both CT and resolutions are kept constant, and the performance of image sharpening and histogram equalization is evaluated.

#### 4.1 Model 1: Suitable CT Identification Using Default SSD

SSD can have a good TP for a given CT on a given image. However, for the same image, it can also have FP. Therefore, there is a need to identify a suitable CT by analysing both precision (representing the detection performance of a detector) and recall (representing the accuracy in regards to number of candidate objects in GT).

To compare and evaluate how CT affects the recall rate on the dataset, scoreR is computed one each image using different CTs. The winner CT is identified based on best scoreR.
The number shown on the curve in the Fig. 4, is the sum of marks each CT gets for entire dataset. The horizontal axis indicates different CT (from 0.1 to 0.9), while the vertical axis represents the score CTs get.

The SSD detector starts detecting objects more often when they become larger in size, i.e. from image 282 onwards. Therefore, there is no winner for images from 0 to 281, so for the further analysis based on recall these images were ignored based on the experiment.

For images 282 onwards, CT = (0.1 || 0.2) has the best recall-based winner. To further analyse this, let’s consider an image 293 in EISATS set 9 where CT = (0.1 || 0.2) is a winning combination. To be more specific, $R_{new}^{0.1} = R_{new}^{0.2} = 0.5$, $R_{new}^{0.4} = R_{new}^{0.5} = \ldots = R_{new}^{0.9} = 0.25$ according to our experiments.

Figure 5 shows an example of image 293 from EISATS set 9 with different SSD detections due to various CTs.
The lower CT chosen, the better is the recall rate demonstrated. This is reasonable, because some cars are detected with a lower confidence. If a lower confidence threshold is chosen it will pass the filter and be shown as a detection in the output. However, for a higher CT than the confidence, it will not pass the filter of CT and consequently cannot be shown in the detection result. That will lead to a lower TP, as well as lower recall.

(2) All the detected cars come with confidence more than 20 percent, therefore CT = 0.1 and CT = 0.2 win in all the images, or at least perform as well as other CT (when no detection is produced).

To compare precision between different CTs scoreP was used. However, unlike scoreR, the scoreP curve is nonlinear (see Fig. 6A). Instead of a highest CT or lowest CT, CT of 0.7 using \( P_{\text{new}}^{0.7} \) has best performance.

Fig. 6. Comparison for different CTs on EISA TS set 9 using scorePR as marks. (A).Precision only. (B). Precision and recall.

From our previous experiment we know that lower CT can produce more detections. However, most of the detections at lower CT are FPs. These are due to the detections of buses, boats, aeroplanes, chairs and trains. Neither of these are present in GT, hence they are FP detections. \( P_{\text{new}}^{CT} \) determines how many of detections out of total detections were TP. This explains the lower score of \( P_{\text{new}}^{CT} \) for lower CTs.

The reason that \( P_{\text{new}}^{0.7} \) gets higher score in scoreP system than \( P_{\text{new}}^{0.8} \) and \( P_{\text{new}}^{0.9} \) is mainly because some detections come with a medium confidence. For example, in image 381, only one detection with confidence 79.4 percent is produced, no false detections. In such a situation, CT = 0.1 ~ 0.7 lead to precision of 100 percent, but CT = 0.8 ~ 0.9 lead to 0 percent. This is not a common scene among the whole dataset, so the value of CT = 0.8 and 0.9 are lower but relatively near to CT = 0.7.

To find the best fit for both recall and precision, both are plotted together (see Fig. 6B). Based on it, both the curves meet at CT = 0.3. This initially indicates CT = 0.3 as a candidate for a trade-off of recall and precision. This concludes hypothesis 1 and is used in following models.

4.2 Model 2: Suitable Resolution Using Default SSD and CT = 0.3

Each greyscale image in the EISATS set 9 dataset is of resolution \([W \times H] = [640 \times 480]\) pixels. Where, W indicates horizontal count of pixels, while H stands indicates...
vertical count. The default SSD model resizes input image to \([W \times H] = [300 \times 300]\) pixels.

We scale up the image up to 1.25 times i.e. \([800 \times 600]\) pixels, and then down to 0.75 times i.e. \([480 \times 360]\) pixels, and 0.5 times i.e. \([320 \times 240]\) pixels.

**Fig. 7.** ScorePR values for different input image resolutions and \(CT = 0.3\) on EISATS set 9. Top: Precision score. Bottom: Recall score.

Based on Fig. 7 where comparison is only about original or scaled down resolution, the resolution \([W \times H] = [640 \times 480]\) pixels gains highest number of peaks most frequently, therefore it is a clear winner. Similarly, when the resolution is reduced, SSD recall’s performance degrades accordingly. Additionally, we found that the method used to change the scale of image can also impact the performance of SSD. In our analysis, we compared nearest neighbour interpolation (NNI) with bilinear interpolation, NNI with area, bi-cubic interpolation and with Lanczos interpolation. We found that the method used for scaling up, can positively improve SSD’s performance. Based on which we were able to identify that with NNI method, resolution \([800 \times 600]\) pixels outperformed the image’s default resolution i.e. \([640 \times 480]\) pixels.

4.3 Model 3: Preprocessing Comparison on SSD Models with \(CT = 0.3\) and Resolution = \([800 \times 600]\) Pixels

Figure 8 illustrates a comparison based on scorePR on complete dataset without pre-processing, with image sharpening only, with histogram equalisation only, with image sharpening as well as histogram equalisation. Based on this, it is clear that the default model without any pre-processing relatively has the worst performance. With pre-histogram equalisation only, SSD gives the best recall score. With both image sharpening and histogram equalisation as pre-processing, SSD gives the best result for precision. Based on this it can be concluded that there is not a single winner. However, pre-processing does improve the overall outcome of SSD. And, same can be expected for other more recent object detectors.
5 Conclusions

CT is a basic factor which could impact the result of SSD detection. In our experiments on EISATS set 9, we found that CT = 0.3 outperformed others when evaluations was performed using the novel scorePR system. Furthermore, based on further experiments we deduced that the resolution of images is crucial to the SSD. Scaling up or scaling down the resolution of images could change the performance of SSD detection results.

We further explored the impact of visual image quality upon the accuracy of SSD using scorePR. Specifically, improving image quality by histogram equalisation and image sharpening could effectively improve the SSD accuracy. Based on these findings, a few new models of SSD were proposed, including iSSDv1 and iSSDv2.

We also believe that accuracy of more recent detectors can be improved by improving image quality as well as image resolution. Furthermore, it is important to maintain the scale of object in both directions when scaling up or down. The performance of proposed models can be evaluated with scorePR as well.

References

1. ‘WHO—Global status report on road safety 2018. WHO (2018). https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/. Accessed 18 Aug 2019
2. Khan, W., Morris, J.: Safety of stereo driver assistance systems. In: 2012 IEEE Intelligent Vehicles Symposium, Alcala de Henares, pp. 469–475 (2012). https://doi.org/10.1109/IVS.2012.6232188
3. Khan, W., Morris, J., Klette, R.: Stereo accuracy for collision avoidance. In: 2009 24th International Conference Image and Vision Computing New Zealand, Wellington, pp. 67–72 (2009). https://doi.org/10.1109/IVCNZ.2009.5378358
4. Wu, Z., Thangali, A., Sclaroff, S., Betke, M.: Coupling detection and data association for multiple object tracking. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, pp. 1948–1955 (2012). https://doi.org/10.1109/CVPR.2012.6247896
5. Hermann, S., Morales, S., Klette, R.: Half-resolution semi-global stereo matching. In: 2011 IEEE Intelligent Vehicles Symposium (IV), pp. 201–206, June 2011. https://doi.org/10.1109/IVS.2011.5940427

6. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE CVPR, Providence, RI, pp. 3354–3361, June 2012. https://doi.org/10.1109/CVPR.2012.6248074

7. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE CVPR 2001, vol. 1, p. I–I, December 2001. https://doi.org/10.1109/CVPR.2001.990517

8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE CVPR 2005, San Diego, CA, USA, vol. 1, pp. 886–893 (2005). https://doi.org/10.1109/CVPR.2005.177

9. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: 2008 IEEE CVPR 2008, pp. 1–8 (2008). https://doi.org/10.1109/CVPR.2008.4587597

10. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1627–1645 (2010). https://doi.org/10.1109/TPAMI.2009.167

11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)

12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE CVPR, Columbus, OH, USA, pp. 580–587, June 2014. https://doi.org/10.1109/CVPR.2014.81

13. Girshick, R.: Fast R-CNN, pp. 1440–1448 (2015). Accessed 30 July 2019

14. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 91–99. Curran Associates, Inc. (2015)

15. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2

16. Olson, D.L., Delen, D.: Advanced Data Mining Techniques. Springer, Berlin (2008)

17. Khan, W., Suaste, V., Caudillo, D., Klette, R.: Belief propagation stereo matching compared to iSGM on binocular or trinocular video data. In: 2013 IEEE Intelligent Vehicles Symposium (IV), pp. 791–796, June 2013. https://doi.org/10.1109/IVS.2013.6629563