Bicycle Detection from Top View Perspective in Surveillance System using Convolutional Neural Network

Sanal Darshid Ramkumar[1]

Abstract: Bicycle detection and tracking with cameras from top view perspectives using deep learning is a highly active research area for video surveillance and automatic ticket generation in Advanced Public Transportation System (APTS). People detection using conventional cameras has received massive attention for video surveillance inside public transportation systems but inattentive towards bicycle detection. Experimentation is performed on You Only Look Once (YOLO), Faster Regional-Convolutional Neural Network (Faster R-CNN) and Single Shot Multibox Detector (SSD). Due to the sparse availability of dataset for this work, a customized dataset was recorded in the Media Computing lab, Junior Professorship of Media Computing, TU Chemnitz, Germany. True positive and false positive analysis are done to find the best case solutions to reduce the problems affecting the performance of deep learning models due to occlusions and view point variations. This paper provides best case solution for bicycle detection from a top view perspective and has achieved mean average precision of 92 %.

Keywords: YOLO, Faster R-CNN, SSD

1 Introduction

Humans and computers have inborn conditions suited for performing different types of tasks. For example, calculating the cube root of a large number is very easy for a computer, but difficult for humans. The human visual system is fast and accurate, allowing us to perform complex tasks like driving with little conscious thought [Fe10]. On the other hand, a task such as recognizing the objects in an image is easy for humans but has traditionally been very difficult for an automated learning algorithm. The term object detection refers to object localization and classification [Fe10]. Generic object detection methods are based on region proposal and regression/ classification [Zh19].

Closed Circuit Television (CCTV)/ Internet Protocol (IP) cameras and network infrastructure have become cheaper and more affordable. Thus providing a new surveillance technology that applies to a wide range of end-users in retail sectors, schools, homes, office, industrial/ transportation systems, and government sectors. Over the years, with the advances of deep learning, more concretely Convolutional Neural Networks (CNN) [HS06], image recognition and object detection have been progressing at a rapid pace. Most of this progress is not just the result of powerful hardware, bigger datasets and models, but mainly a consequence

[1] Chemnitz University of Technology, Junior Professorship Media Computing, Straße der Nationen 62, 09111 Chemnitz, Germany sanal-darshid.ramkumar@informatik.tu-chemnitz.de

of new ideas, algorithms and improved network architectures. The real-world detection and counting of a bicycle are hard, which involves locating bicycle in different scenes irrespective of orientation, scale, environment and type of view.

The paper is structured in the following sections as Related work 2, Dataset preparation 3, Experimental setup 4, Implementation 5 and Results 6.

2 Related Work

This section provides overview of object detection in public transportation systems, as well as the challenges for bicycle detection.

An APTS is an application of Intelligent Transportation Systems (ITS) [Zh18] which aims to provide services relating to different modes of transport and traffic management resulting in a safer, more coordinated and smarter use of transport networks. Research works are focused on fields like calling for emergency services when an accident occurs, using cameras to enforce traffic laws or signs that mark speed limit changes depending on conditions, usage of CCTV for surveillance. The directive of the European Union (EU) 2010/40, on July 7, 2010, defined ITS [Li18] as a system that applies information and communication technologies in the field of road transport, infrastructure, vehicle users, traffic and mobility management [Le89]. High demand is seen in today's surveillance systems in APTS for detection and counting of people and bicycle for surveillance, automatic ticket generation and to display the availability of free slots for parking bicycles in public transport systems such as bus and train [VM19].

Many research works are focused on people detection but countable in terms of bicycle detection for applications in public transportation systems. Cameras deployed in top view perspective to monitor a large scene with limited resources for people and bicycle detection are crucial. Nevertheless, the problem has received only limited attention, and state of the art lags behind bicycle detection in conventional top view perspectives. A research work based on traditional machine learning [Sh15] is available but limited to viewpoint variation. Research work in APTS for bicycle and motorcycle detection discussed on effect of different climate, clearly this paper focuses on traffic monitoring from the data gathered from CCTV [De21]. Another work towards safety of cyclist focuses on orientation using deep learning in the field of autonomous vehicles [GMC20]. The previous research works lags in identifying effects of radial distortion, elevated viewpoint and deformation from the top view, extensive variability of bicycle appearance from top view leads to difficulty in obtaining features for accurate detection. These features are sensitive to noise, occlusion of multiple bicycles, illumination changes and viewpoint variations. Multiple cameras deployed in the top view perspective might provide a better solution for bicycle detection.

In this work experimentations are done on YOLOv2 using Darknet and PyTorch [Sa18], [Op18], YOLOv3 using Darknet and PyTorch [RF18], [Op18], SSD using Inception v2 [Li16], Faster R-CNN using Inception v2, ResNet-50 and ResNet-101[Re16] architecture.

3 Dataset Preparation

For surveillance systems, privacy and comfort are key factors to be considered, as they should not impose pressure on anyone. This research work demands a dataset from a top view perspective for applications in indoor and outdoor surveillance systems for bicycle detection and tracking. As there is a shortage of publicly available dataset, a customised dataset was created. This customised dataset is created concerning laboratory and bicycle parking scenarios where humans take bicycle along with them to the respective areas. The customised dataset (DS1 and DS2) are recorded in the Media Computing lab, Junior Professorship of Media Computing, TU Chemnitz, Germany. Visual information is retrieved by recording videos in High Definition (HD) resolution from 10 optical smart embedded stereo sensors (S2000, Intenta GmbH). To further enhance the quantity of the dataset two more datasets (DS3 and DS4) recorded using mobile camera were included. Data augmentation like flipping, scaling and cropping are performed on DS3 and DS4.
Videos were converted to images and labelled using the Computer Vision Annotation Tool (CVAT) [BM]. Sample images available in the dataset is depicted in figure 1. Key facts and typical differences are presented in table 1.

- DS1 - Lab dataset consisting of 207 videos recorded using a wide-angle smart stereo sensor (S2000, Intenta GmbH) mounted in top view perspective with bicycle and human in an indoor scene. Sample images are shown in the first row of figure 1
- DS2 - Lab dataset consisting of 232 videos recorded using a wide-angle smart stereo sensor (S2000, Intenta GmbH) mounted in top view perspective with bicycle and human in an outdoor scene. Sample images are shown in the second row of figure 1
- DS3 - An indoor dataset consisting of 15 videos recorded using a mobile camera in top view perspective with bicycle and human in a bicycle garage. Sample images are shown in the third row of figure 1
- DS4 -An outdoor dataset consisting of 15 videos recorded using a mobile camera in top view perspective with bicycle and human near the TU Chemnitz university area. Sample images are shown in the fourth row of figure 1

Dataset	**Application**	**Field of view (degree)**	**Position of Camera from ground (meter)**	**Total Camera Deployed**	**Videos Generated**	**Average Video length**
1	Indoor Lab Scene	97	5	10	207	20 sec
2	Outdoor Lab Scene	97	5	1	232	31 sec
3	Indoor Bicycle garage	64	6.5	1	15	50 sec
4	Outdoor University area	64	6.5	1	15	50 sec

Tab. 1: Dataset Overview

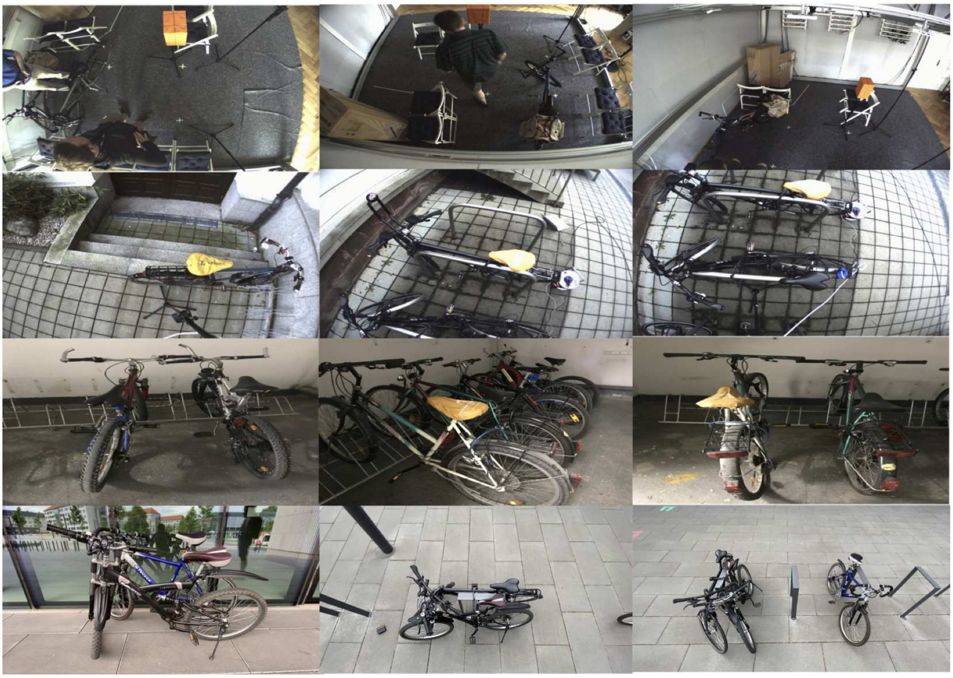

Fig. 1: The images in the first row shows DS1 dataset, the second row shows the DS2 dataset, the third row shows the DS3 dataset and the fourth row shows DS4 dataset

Combinations of dataset for training

- The dataset is split into train (80%), validation (10%) and test (10%)
- For applications related to indoor surveillance system, DS1 and DS3 were combined
- For applications related to outdoor surveillance system, DS2 and DS4 were combined
- For general purpose surveillance system application, all four dataset were combined

4 Experimental Setup

For this research work the following hardware and software configurations are used.

Hardware Configuration Intel Core i7 with 8GB RAM, CPU (4 cores at 3.60 GHz) with Ubuntu Bionic Beaver as operating system is used. NVIDIA GeForce RTX 2070 SUPER 8GB is the GPU used.
Software Configuration Python 3 with cuda 10.0 tool kit for GPU accelaration and cuDNN 7.5 libraries are used. TensorFlow, Darknet and PyTorch are the framework used.

Dependencies such as numpy 1.14.2, scipy 1.0.0, python-openCV, matplotlib 2.2.0 and cython 0.29.2 are used.

5 Implementation

This section describes in detail the implementation of all the eight models introduced in section 2. Furthermore, this facilitates to benchmark the obtained results in bicycle detection.

5.1 Configuration setup

The configuration details of the eight models, YOLOv2 (Darknet and PyTorch), YOLOv3 (Darknet and PyTorch), SSD (Inception v2), Faster R-CNN (Inception v2, ResNet-50 and ResNet-101) are shown in the table 2. From the table 2, the Faster R-CNN models used a batch size of 1 and this restriction was due to the hardware configuration limiting the conversion of tensor shape. Table 3 shows the details of the number of training and validation images used for the different dataset combinations used in this research work. For this research work, the default settings of hyperparameters worked well.

S.no	Model	Architecture	(Height,Width, Channel)	Batch size	Epochs
1	Faster R-CNN [Re16]	Inception v2	(600,1024,3)	1	200
2	Faster R-CNN [Re16]	ResNet-50	(600,1024,3)	1	200
3	Faster R-CNN [Re16]	ResNet-101	(600,1024,3)	1	200
4	SSD [Li16]	Inception v2	(300,300,3)	24	200
5	YOLO v2 [Sa18]	Darknet-19	(416,416,3)	64	250
6	YOLO v2 [Op18]	Darknet-19 PyTorch	(416,416,3)	64	80
7	YOLO v3 [RF18]	Darknet-53	(416,416,3)	64	250
8	YOLO v3 [Op18]	Darknet-53 PyTorch	(416,416,3)	64	80

Tab. 2: Configuration

Dataset	Training Images	Validation Images
DS1 and DS3	2983	200
DS2 and DS4	3647	300
DS1,2,3,4	6630	600

Tab. 3: Overview of Training and Validation

6 Results and discussion

Analysis of Mean Average Precision (mAP)
The goal of an object detector is to predict the location of a given class in an image or video with a high confidence value. This is done by placing bounding boxes to identify the position of objects. This is done by a set of three attributes: 1) Object class, 2) Corresponding bounding box and 3) Confidence score usually given by a value between 0 and 1. The evaluations are done based on :

- A set of ground-truth bounding boxes representing the rectangular areas of an image containing objects of the class to be detected
- A set of detections predicted by a model, each one consisting of a bounding box, a class, and a confidence value

Average precision (AP) is a popular metric in measuring the accuracy of object detectors. It computes the average value of the precision over the recall r interval from $r = 0$ to $r = 1$. The mean average precision (mAP) is the average of AP, which quantifies how good the model is at performing the query, results are shown in figure 4. In this work, true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) based on Intersection of union (IoU) are experimented and discussed in this section for analyzing the limitations from the top view perspective. In this research work, mAP is obtained for different combinations of dataset type as shown in section 3.

- True Positive (TP): A correct detection. Detection with IoU >= threshold
- False Positive (FP): A wrong detection. Detection with IoU < threshold
- False Negative (FN): A ground truth not detected
- The threshold for evaluation is set as 70 %

TP, FP and FN analysis for limitation in bicycle detection are shown in figures 2 and 3 from the results of the best model (8, see table 2) obtained for this work. The number of TP, FN and ground truth are discussed. The green bounding box represent ground truth, red bounding box represent the detected true positives and yellow bounding box represents false negatives. This analysis allows to determine the limitations such as view point variation, occlusion and detecting the exact number of bicycle present and for the evaluation of the mAP. When observed upto two bicycles there is no occlusion problem when multiple bicycles overlap (see figure 3). Whereas, for more than two bicycles, occlusions result in false negatives (see figure 2).

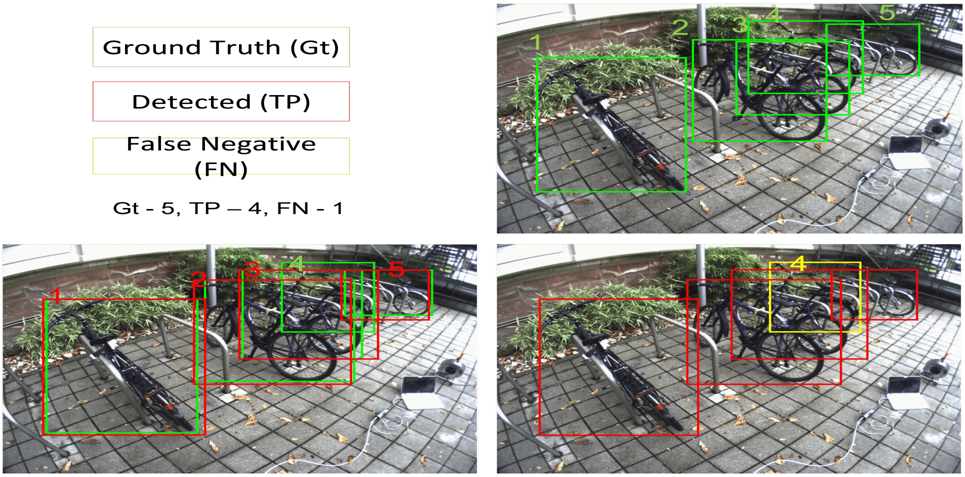

Fig. 2: TP, FP and FN analysis for limitation in bicycle detection. Here five ground truths present, but only four are detected. The ground truth which is not detected is a false negatives (FN) and it is shown in the last image in the second column with yellow bounding box.

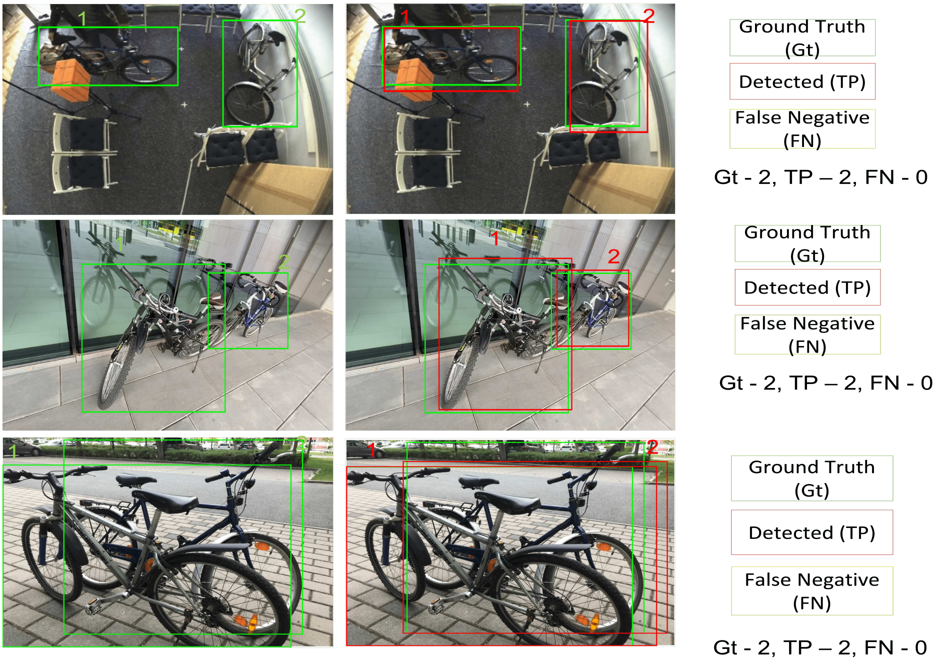

Fig. 3: TP, FP and FN analysis for limitation in bicycle detection.

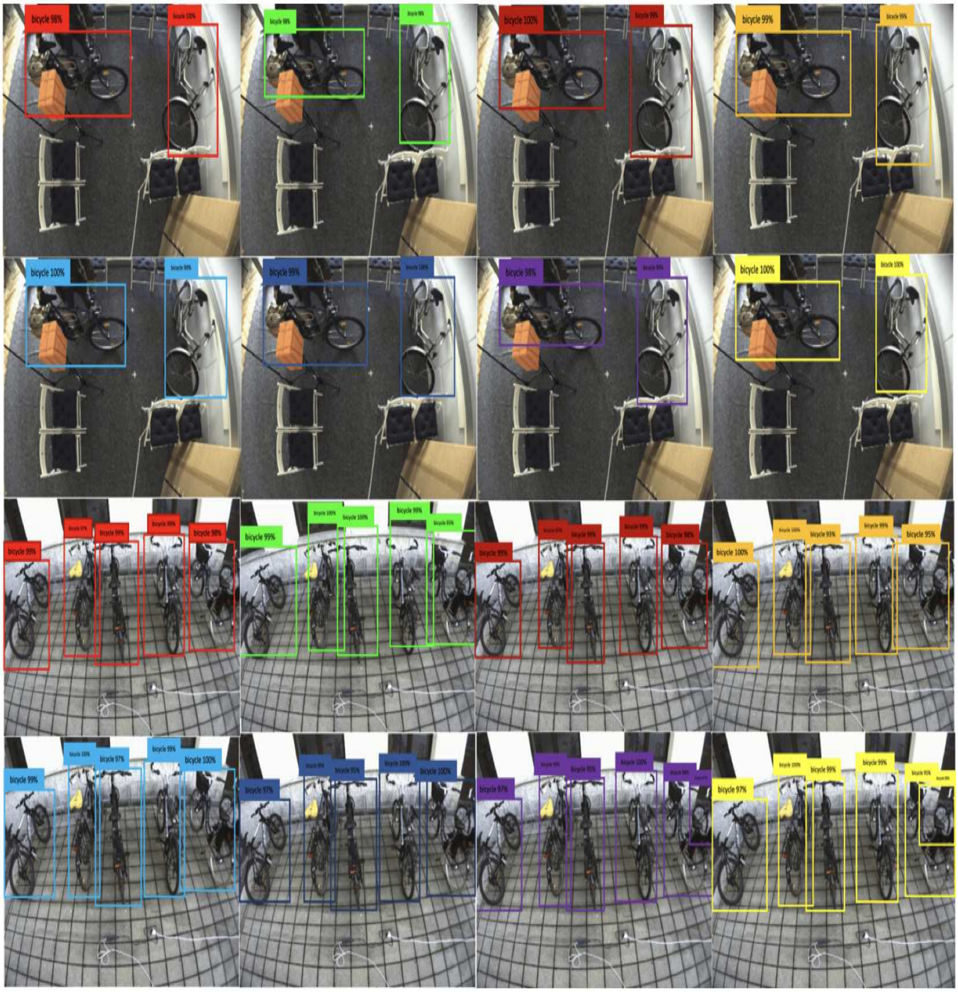

Fig. 4: Sample output from all eight models trained for combined dataset DS1, DS2, DS3 and DS4. The figure has 16 images. The eight models has unique colour for the predicted bounding box here. The first two rows are results from the Indoor dataset DS1 and DS3 and the last two rows are results from the outdoor dataset DS2 and DS4. The outputs of the models represented here are as follows with the first image in the first row as number 1 to the last image in the second row as 8 for the Indoor dataset, the same numbering is repeated for the last two rows of the image for Outdoor dataset. No.1 depicts FR-CNN Inception v2, No.2 depicts FR-CNN ResNet-50, No.3 FR-CNN ResNet-101, No.4 depicts SSD Inception v2, No.5 depicts YOLOv2 Darknet, No.6 depicts YOLOv3 Darknet, No7. depicts YOLOv2 PyTorch and No.8 depicts YOLOv3 PyTorch.

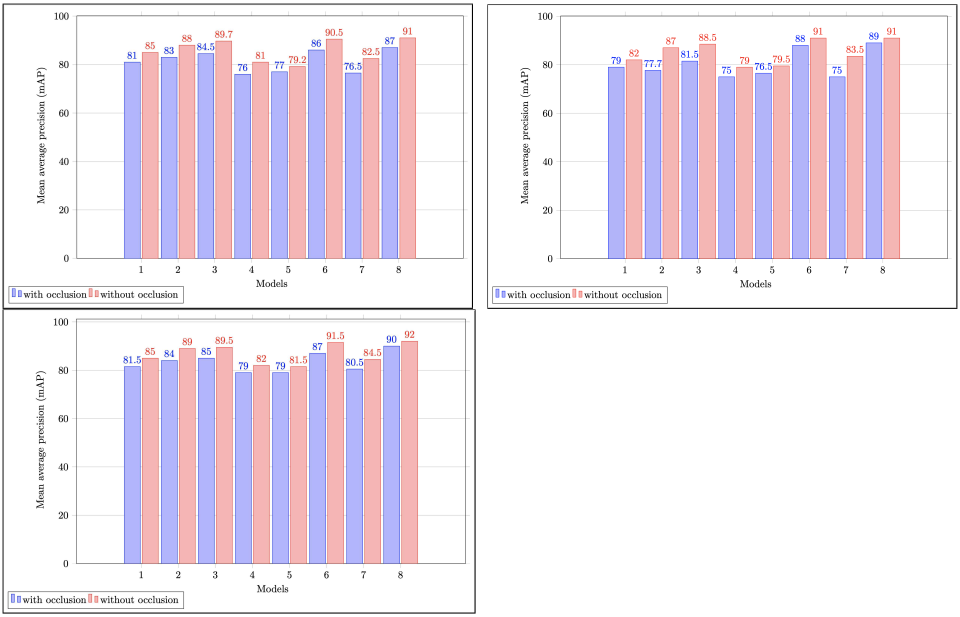

Fig. 5: The figure present here has three graphs which explain the Mean Average Precision obtained in y-axis and on x-axis all eight models shown in table 2. The first row shows results from dataset combination DS1 and DS3. The second row shows results from dataset combination DS2 and DS4. The third row shows results from combining all four dataset DS1,DS2, DS3 and DS4. The results are obtained based on with and without occlusions of bicycle.

From the fig 5, it is to be noted that, for this research work the number of bicycle present in a scene is limited to 10. Thus when experimenting with this dataset, good results were obtained. The dataset used in this research work was recorded from single view point but from different multiple camera deployed in a scene. Thus enabling to reduce problems occurring from view point variations. The mAP variations between with occlusion and without occlusion are less because the research work is focused more towards data driven approach, rather than model driven. This enables in achieving best results. It is clearly seen that model 8 out performs all other models.

7 Conclusion and Future Work

This work deals with bicycle detection from a top view perspective for surveillance application in bicycle parking area, inside public transportation's and near university area where there is more demand in everyday bicycle usage. Due to the scarcity of publicly available dataset, four customised datasets were prepared. They were combined into three categories and were used to benchmark the results on eight deep learning models. Evaluations are performed using mAP, TP, FP and FN analysis. The limitations were due to the single view perspective analysis of bicycle detection from the top view. One of the tedious work here was gathering the data and annotating multiple overlapped bicycles. In general overlapping of objects effect the performance of a deep learning model. Thus multiple overlapping of bicycle with bicycle in different view point may reduce the performance of detection. As the research work is well organized and data is gathered perfect, we have achieved better results.

During this research work, an occlusion problem occurred when one bicycle is completely hidden behind another bicycle from a top view perspective. This limitation can be encountered through quantifying the existing dataset by including synthetic datasets. Current work is under progress to overcome this limitation. This work focuses on creating a synthetic dataset using a game engine called Unreal Engine 4 (UE4), where the synthetic dataset represents scenes inside trains and trams used in Germany with humans and bicycle. This work focuses on bicycle detection from a multi-view perspective, where the real world and synthetic data fusion may provide better results. Furthermore, to reduce the effort and time involved in tedious process of manual labelling in supervised technique, the game engine generates ground truth for the synthetic dataset by incorporating required plugins and scripts.

Bibliography

[BM] Boris Sekachev, A. Z.; Manovich, N./, URL: https://github.com/openvinotoolkit/cvat.

[De21] Dequito, C.; Dichaves, I.; Juan, R.; Minaga, M.; Ilao, J.; Cordel, I. M.; Del Gallego, N.: Vision-based bicycle and motorcycle detection using a YOLO-based Network. In: Journal of Physics: Conference Series. Bd. 1922. 1, IOP Publishing, S. 012003, 2021.

[Fe10] Felzenszwalb, P.; Girshick, R.; Mcallester, D.; Ramanan, D.: Object Detection with Discriminatively Trained Part-Based Models. IEEE transactions on pattern analysis and machine intelligence 32/, S. 1627–45, Sep. 2010, URL: https://ieeexplore.ieee.org/document/5255236.

[GMC20] Garcia-Venegas, M.; Mercado-Ravell, D. A.; Carballo-Monsivais, C. A.: On the safety of vulnerable road users by cyclist orientation detection using Deep Learning. arXiv preprint arXiv:2004.11909/, 2020.

[HS06] Hinton, G.; Salakhutdinov, R.: Reducing the Dimensionality of Data with Neural Networks. Science (New York, N.Y.) 313/, S. 504–7, Aug. 2006, URL: https://www.researchgate.net/publication/6912170_Reducing_the_Dimensionality_of_Data_with_Neural_Networks.

[Le89] Lecun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; Jackel, L.: Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation 1/, S. 541–551, Dez. 1989, URL: https://ieeexplore.ieee.org/document/6795724.

[Li16] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.: SSD: Single Shot MultiBox Detector. In. Bd. 9905, S. 21–37, Okt. 2016, ISBN: 978-3-319-46447-3, URL: https://arxiv.org/abs/1512.02325.

[Li18] Liu, X.-Y.; Ding, Z.; Borst, S.; Walid, A.: Deep Reinforcement Learning for Intelligent Transportation Systems, Dez. 2018, URL: https://arxiv.org/abs/1812.00979.

[Op18] Ophoff, T.: Lightnet: Building Blocks to Recreate Darknet Networks in Pytorch./ , 2018.

[Re16] Ren, S.; He, K.; Girshick, R.; Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In. S. 1–10, Jan. 2016, URL: https://ieeexplore.ieee.org/document/7485869.

[RF18] Redmon, J.; Farhadi, A.: YOLOv3: An Incremental Improvement./, Apr. 2018.

[Sa18] Sang, J.; Wu, Z.; Guo, P.; Hu, H.; Xiang, H.; Zhang, Q.; Cai, B.: An improved YOLOv2 for vehicle detection. Sensors 18/, S. 4272, Dez. 2018, URL: https://www.researchgate.net/publication/329409946_An_improved_YOLOv2_for_vehicle_detection.

[Sh15] Shahraki, F. F.; Yazdanpanah, A. P.; Regentova, E. E.; Muthukumar, V.: Bicycle Detection Using HOG, HSC and MLBP. In (Bebis, G.; Boyle, R.; Parvin, B.; Koracin, D.; Pavlidis, I. T.; Feris, R. S.; McGraw, T.; Elendt, M.; Kopper, R.; Ragan, E. D.; Ye, Z.; Weber, G. H., Hrsg.): Advances in Visual Computing - 11th International Symposium, ISVC 2015, Las Vegas, NV, USA, December 14-16, 2015, Proceedings, Part II. Bd. 9475. Lecture Notes in Computer Science, Springer, S. 554–562, 2015, URL: https://doi.org/10.1007/978-3-319-27863-6%5C_51.

[VM19] Veres, M.; Moussa, M.: Deep Learning for Intelligent Transportation Systems: A Survey of Emerging Trends. IEEE Transactions on Intelligent Transportation Systems PP/, S. 1–17, Juli 2019, URL: https://ieeexplore.ieee.org/document/8771378.

[Zh18] Zhang, Z.; Wang, Y.; Chen, P.; Yu, G.: Intelligent Transportation Systems (ITS). In. S. 444–454, Jan. 2018, URL: https://www.researchgate.net/publication/326487592_Intelligent_Transportation_Systems_ITS.

[Zh19] Zhao, Z.-Q.; Zheng, P.; Xu, S.-T.; Wu, X.: Object Detection With Deep Learning: A Review. IEEE Transactions on Neural Networks and Learning Systems PP/, S. 1–21, Jan. 2019, URL: https://ieeexplore.ieee.org/document/8627998.