Proceedings of 9th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2021 Month: March Volume: 9

Comparative study of CCTV based Vehicle Identification and Classification Models during Adverse Conditions in Pokhara

Biplav Sharma Regmi [a], Ramesh Thapa [b], Biplove Pokhrel [c]

[a, b, c] *Department of Electronics and Computer Engineering, Paschimanchal Campus*
Corresponding Email: [a] biplav01regmi@gmail.com

Abstract
Vehicle detection and counting mechanisms using existing security CCTV cameras can be very useful to estimate traffic flow prediction and anomaly detection for better road planning. With the rapid enhancement in the computational ability, applying deep learning algorithms has been a increasing area of research for vehicle detection and classification. The targeted problem of this research is formulated as vehicle detection and classification in different adverse conditions. This paper compares the state-of-art method for vehicle detection and classification in custom Nepali dataset, in which categories of vehicle and rules of annotation are specified. The custom dataset consists of significant number of main frames from a high-resolution (2304×1296) CCTV camera captured at 24fps, which is pointed towards Birauta-Syanjha connecting road and mounted in front of Chorepatan police check-post. This dataset consists of keyframes of challenging foggy morning conditions, crowded peak daytime traffic and nighttime key-frames. The experimental result shows that among different existing single stage and double stage state-of-art classifiers, pre-trained YOLO-v5 model with CSPDarkNet as backbone architecture on COCO dataset outperforms other models. The dataset presented here might be further used by other researchers as additional training data or challenging dataset to purpose a novel vehicle detection and classification system at adverse climatic conditions.

Keywords
Vehicle detection, Vehicle Classification, CNN, Traffic analysis, Surveillance data

1. Introduction

Vehicle counting is still based on classical approach of manual interpretation or by using costly sensors. Traditional image processing method and sensor methods posses low precision and high installation cost. In recent trends, the use of security surveillance camera in roadside area has been increased. These video feeds provides a base for analysing traffic. Despite, the broad deployment of surveillance camera, it is still not used as a way to capture and analyze traffic data for developing Intelligent Transportation System(ITS). Thus, it would be a helpful job to switch camera use from current manual control only to automatic monitoring [1]. Advanced, accurate and fast methods can lead to the implementation of vision based vehicle tracking.

Existing state-of-art vehicle detection method are not so robust in harsh environmental conditions and night-time conditions and exhibits below the par level of performance. There are several restriction for vision based traffic analysis. The detection is significantly affected by restrictions on camera functionality because of unfavorable climatic conditions such as heavy fog, rain, and snow. A wide variation in image frames such as occlusions, scale, view point, illumination, deformation and background cluster is encountered which suppresses the performance of the models. Foggy weather is a major challenge in context of Pokhara. Also, the night time vehicle detection is a another challenge for the implementation of ITS. Using the transfer learning approach helps to improve efficiency of advanced state-of-the-art techniques trained on generalized dataset but transfer learning method can be used with a new dataset for training, to increase the scene specific precision with a pre-trained model. But this requires preparation of new data to train and test the model. Data annotation, itself is a tedious job which requires a huge effort and time.

This research contributes a new Nepali traffic dataset of roadway vehicle images at foggy morning, peak

day-time and night time which are challenging conditions for vehicle detection in context of Pokhara. Vehicles are classified in 4 main categories: Motorcycle, Car, Bus and Truck, which are major contributing categories for parametric estimation in roadway traffic analysis. In addition, the performance of representing state-of-art models on this new dataset are compared. In this study, for comparing the performance of the model, time taken by the model to make inference for a frame is considered. This is essential for determining the feasibility of model to make real-time detection. The metrics considered for this comparison is mean Average Precision(mAP). A score is returned by mAP by contrasting the bounding box of ground truth with the box detected. The higher the score, the more accurate the model is in its detection. This research illustrates that YOLO-v5 model performs well for the vehicle detection and classification task for the traffic scenes in this dataset.

The remainder of our article is arranged as follows. Related works to this study are described in Section 2. Section 3 introduces our approaches for pre-processing the video frames, annotation rules, dataset preparation and comparison of performance of existing deep-learning model on our dataset. The findings of the analysis carried out are discussed in Section 4 and the article is finally concluded in Section 5.

2. Related Works

With the increased interaction in the fields of computer graphics and computer vision, a major shift came about at the end of the 1990s. This included rendering based on images, image morphing, interpolation of views, panoramic image stitching and early rendering of light fields[2]. Since the emergence of the very first research on object detection, image classification systems have become a prospective high valued research field. Vehicle identification and classification has always been a focused area to manage and control traffic issues, which can be done with greater efficiency from video surveillance as compared to other methods.

However, academic researches on Vehicle Tracking System have elated in significant manner over past decades, more research is required to be applicable in the real world situations. That is to say implementing such general research with integrating specific attributes related to the distinct purposes. Despite that, considering the nature of the deep learning architectures available to detect and classify the vehicles, it is a hard task to refine the Moving Object Detection to specific disciplines.

RCNN-based approaches have become mainstream for object detection with the introduction of the CNN network. They can be divided into two classes: the network of two stages and the network of one stage. The crucial distinction among them is the proposal for region. Single stage detectors extract one time function and proposes final regression layer area like You Only Look Once YOLO[3] and Single Shot Detector SSD[4]. There is accuracy precision trade off between single stage and double stage detector. The single stage detector is fast but less precise, while the two-stage detector is more precise but slow.

Ross Girshick et al. implemented one of the core deep learning approaches to object detection[5]. This architecture uses a network of regional proposals in which the suggested bounding box is supplied from an external mechanism such as selective search and fed into a neural convolution network. To classify whether or not an object of interest is present in the bounding box, a support vector machine uses the triggered features detected by the CNN and bounding blocks. Ultimately, a regression layer tightens the bounding box around object. Ross Girshick et all. Improved RCNN with Fast RCNN[6] by incorporating RoiPool. This enhancement increases inference by eliminating the need to make more than one forward pass through the network for each image. In addition, the Soft Max layer and Regression Network are integrated with CNN in the form of the external SVM classifier and Regression model. A final improvement was made on this architecture known as Faster RCNN[7] which uses the insight discovered from features captured by Selective Search(SS). A good improvement in training and testing speeds and mAP ratings is shown by Faster RCNN. Faster RCNN can be trained end to end, most notably. Jifeng Dai et al. suggested a fully convolution network (RFCN) based region that is fully converted with almost all computation shared on the entire image[8]. An advanced Mask RCNN [9] based on RFCN was suggested to detect objects in an image while generating a high-quality segmentation mask for each instance at the same time. Unlike to Faster RCNN, Mask RCNN adds mask branch for pixel to pixel level instance Segmentation.

Traffic Surveillance Research discussed hybrid methods on the calibrated and uncalibrated camera.

This research utilizes Mask R-CNN's Vehicle Detection and Instance Segmentation[10]. Similarly, Aleksandr Fedorov et al. presents their study which purposes to address the issue of estimating traffic flow on the basis of data from a video surveillance camera[11]. They used the Faster R-CNN two-stage detector together with a SORT tracker. Several modifications as focal loss, adaptive feature pooling, additional mask branch, and anchor optimization have improved the baseline performance of the Faster R-CNN in this study. Yaoming Zhang et al. proposed a YOLOv3 algorithm that can effectively detect and track vehicles in real time, increase the precision of video feature extraction and also significantly improve the speed of detection[12]. However these methods are generalised for all type of weather and day-light conditions.

With the concepts of a block-wise context update mechanism, Fei Liu designed an integrated real-time vehicle counting system to reduce the amount of calculation needed and to improve the efficiency of vehicle detection, separate algorithms for day and night, and free and congested traffic flows to improve system robustness and virtual loop or virtual line detection and adaptive line detection[13]. A different method is used to estimate atmospheric illumination and transmissivity of the vehicle detection structure with a pair of encoders and decoders and to create the defogging image first for the foggy environment in [14].The study in [15] used video data from the Naturalistic Driving Study (NDS) and used multiple promising Deep Learning techniques, including Deep Neural Network (DNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Convolution Neural Network (CNN) on a data collection consisting of three weather conditions, including clear, distant fog and near fog.

For night time vehicle detection, potential region of interest of vehicles is identified and validated in [16] by extracting the composite feature of the vehicle's taillight pair and shadow at the bottom of the vehicle based on a wide set of images in different complex environments. Also, a machine learning-based approach proposed in [17] is to classify whether the bright blobs are headlights, taillights, or other illuminating objects and vehicle tracking with occlusion handling is implemented in different traffic circumstances to refine incorrect identification.

Sudha et al. recommends an advanced deep learning approach entitled enhanced You Only Look Once v3 and enhanced visual background extractor technique to identify multi-type and multiple vehicles in a video sequence while tracking using a hybrid Kalman filtering algorithm and particle filtering techniques to find the trace of upcoming vehicles[18]. In this research, 30 fps input videos are checked under various weather conditions, such as sunny, snowy, night and fog. The effect of vehicle detection at the adverse environment condition and night time conditions are being researched with different approach.

Meng et al. constructed the new expressway data set for multi-vehicle detection tasks consisting of a large number of high-resolution sample images 1920×1080 captured by Pan-Tilt-Zoom (PTZ) cameras from real-world expressway scenes (including the variety of climatic conditions and visual angles), in which vehicle categories and annotation rules are specified and a correlation-matched algorithm for multi-vehicle tracking is proposed[19]. Hassaballah has also introduced a new benchmark dataset for research on autonomous vehicle applications under adverse weather conditions called DAWN, consisting of real-world photos obtained from various types of adverse weather conditions[20].

3. Methods

Centered on the test speed and precision, this paper compares the efficiency of the leading state-of-the-art single stage classifier and double stage classifier on our new Nepali Traffic Dataset (NTD). We divide the work in two major sub-task: Dataset Preparation and Model Comparison. We identify them in detail in the following pages.

3.1 Dataset Preparation

3.1.1 Dataset

Several studies indicate that natural pictures can yield strong identification results based on deep learning algorithm. To improve the state-of-the-art methods for specific task object identification and classification using CNN networks, a number of datasets like ImageNet and PASCAL VOC were already developed.[21]. We present a novel Nepali Traffic Dataset (NTD) dataset to support further research in the field of vehicle detection and classification in adverse conditions. Different from available dataset, we conduct the following work to construct NTD.

3.1.2 Data Collection

Research is based on a single camera for this job, which tracks one of the center point in the linking road between Kaski and Syanjha. This is a entry point and exit point from Kaski and has two sightseeing spot (Devis Fall and Gupteshwor Cave) located on the either side of the roads increasing the possibility of more traffic incidence at the spot. It can remotely monitor the road section as shown in Figure 2. The camera used in this work provides 24 fps, maintaining 2304×1296 resolution. However, owing to blurring, hardware flaws and spider webs, the captured video stream is not flawless. Since the major weather challenge of Pokhara is foggy and rainy weather, we collected foggy data frames which is mostly recorded during morning period. However rainy conditions were not included as we did-not encounter rainy climate during the data collection period.

Compared to the current dataset for vehicle identification[22, 23], we have not annotated long continuous video clips, as complex and crowded scenes will be highly time consuming. We instead concentrated on short clips covering various traffic circumstances. Three conditions were focused for this study. Two adverse climatic conditions: foggy and night-light conditions are considered. We noted that vehicles can be annotated with high confidence in the foggy and night-light atmosphere as well. Peak day time condition is also selected to ensure that instances at all time of a day are included which makes our dataset more robust. Also, peak time condition is a key performer for vehicle detection model since it is a computationally challenging scenario containing large no. of vehicle instances in a single frame.

Collected dataset contains clipped video having at-least a instance of a vehicle in the frame. These clipped videos are merged in 3 categories: Morning Foggy, Peak Daytime and Night time. The morning foggy set contains the video clips from 6 A.M. to 8 A.M. at low, medium and high foggy state at different dates. Similarly peak day time data contains video clips from 8 A.M. to 5 P.M containing different day time illumining condition. Lastly, the Night time data contains video clips of traffic recorded between 5 P.M. to 9 P.M. The traffic instances between 9 P.M. to 6 A.M. are not focused deliberately since there were very few instances of vehicle at this time and also the vehicles appearing at this time in camera frame can't be correctly annotated with high confidence.

Figure 1: CVAT UI

3.1.3 Vehicle Annotation

We annotated 53601 instances across 7429 frames with rectangles.In the Computer Vision Annotation Tool (CVAT)[24], all related annotation tasks are performed. . Among varieties of annotation tools like LabelImg, VGG annotator tool, COCO annotator tool we selected CVAT for its robustness in data annotation and deployment. CVAT is a docker based annotation tool with advanced annotation functionality like interpolation features. The User Interface (UI) of the CVAT is shown in Figure 1.

Vehicles can be classified into two types with regard to the domestic vehicle standard classification manual: Heavy vehicle (truck and bus) and light vehicle (car and motorcycle). A total of 4668 foggy situation key frames, 1129 peak time key frames and 1695 night time key frames were compiled to build our Custom dataset. The performance of the data set relies on the annotation rule used, which are:

1. Small Target: The farther the object gets, the less vehicle's features are acquired. Small items are also carefully labeled with potential compressed bounding boxes from the very far point of view.

2. Occlusion: CVAT annotation tool has a special feature for tackling occlusions in the scene. This feature enables to keep occluded object bounding box behind the occluding object bounded box.

3. Special Samples : A few special samples with category ambiguity like Tractor, Dozer are kept in the data set but are not labelled as any of the Classes used in this data set.

The cumulative figures for the obtained data are summarized in Table 1.

Figure 2: Representing Image Frame at Different Condition used in Dataset

3.2 Model Comparison

Meanwhile, the performance of representative state-of-art methods for two types of classifier i.e. single stage and double stage are compared for this novel dataset. Two representative model are selected by studying the previous research conducted in this field. In the recent trend YOLOv3, among the single stage classifier and Mask RCNN, among the double stage classifier, are widely used. Recent research result performed on this field suggests YOLOv3 generates inference in real-time which is desirable to apply in some online system like vehicle counting and vehicle speed prediction. YOLOv3 model has also been used in challenging weather conditions showing comparative higher efficiency in drastic weather condition[18]. The latest YOLOv3 architecture is shown in Figure 3.

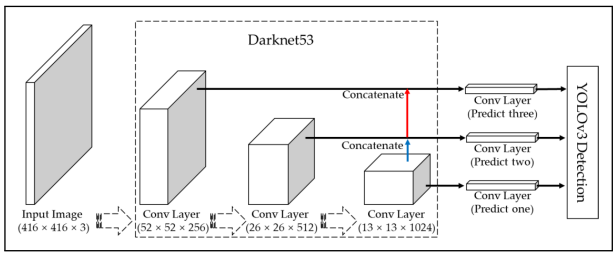

Figure 3: YOLOv3 with Darknet Framework [25]

Mask RCNN, meanwhile, is the leading model in terms of benchmark dataset accuracy. In Mask RCNN, regions are first predicted followed by regression and classification for all predicted regions in second stage. Since, Mask RCNN is most recent two stage detector, it has been used in recent AI challenges for vehicle tracking application [11, 10]. This form of approach is desirable where the results of the prediction, such as anomaly detection, need to be extremely accurate. The architecture of Mask RCNN is shown in Figure 4.

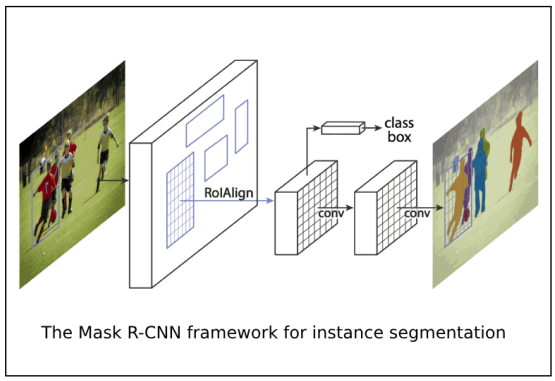

Figure 4: Mask RCNN architecture [9]

There are different caffe deep-learning models available which uses a backbone feature extraction architecture and are trained on benchmark dataset. For this research, Mask RCNN with Inception V2 backbone and YOLOv3 with darknet architecture trained on COCO dataset are selected to compute model's speed-accuracy trade-off among the single stage and double stage methods.

3.3 Evaluation protocol

We compare mean average precision(mAP), at detection IOU threshold 0.5($AP_{0.5}$), to find the model's efficiency as used in several major detection benchmarks[22, 26, 27]. This evaluation protocol is performed for the key frames at 2304×1296 resolution. The confidence threshold for detecting the vehicle is set to 0.5 as an standard threshold metrics to perform the inference.

The inference speed of the models are also compared on varying resolutions of the image. For this, the collected data frames are scaled at other two frame sizes: 720×576 and 3840×2160. The inference speeds at these down-sampled and up-sampled frame sizes are compared with original frame size of the dataset.

4. Results

For preparing our dataset, we select the samples in accordance to the sample balance principle to solve the problem of not having comparative ratio between instances of all categories or physical environment. Since the no. of vehicle instances are very high in day, we collected large no. of frames from morning foggy climatic condition to ensure a balanced form of dataset among different weather conditions. However, the no. of instances of car and motorcycle is much higher in our collected dataset which demonstrates a sub-sampled characteristic of today's real world traffic. The instances of truck and bus are comparatively less in the dataset. Since, they occupy larger area in an image frame, these no. of instances should be enough to extract features among these categories for training the model in future work and research. The overall statistics for the collected data are summarized in Table 1.

Table 1: Distribution of data in Nepali Traffic Dataset (NTD)

	Foggy	Peak Time	Night	Total
No. of Frames	4668	1129	1695	7492
Total Instances	12153	19845	21603	53601
Car	5450	11284	9246	25980
Bus	1640	258	1740	3638
Truck	939	210	830	1979
Motorcycle	4124	8093	9787	22004

Two models selected for the comparison is tested in NVIDIA RTX 1660 Ti GPU on our dataset. The inference time taken by the YOLOv3 model seems to have slight improvement on lower resolutions test frame. This suggests that YOLOv3 model can be used without altering the frame resolution, since the performance is not degraded haphazardly with varying resolution. While, Mask RCNN's performance degraded at rapid rate as the resolution of image is increased. The inference time taken by the Mask RCNN for up-scaled (3840×2160) image frame is 4.5 times than for the down-scaled sample (720×576). This result suggests to down-sample the image in lower resolution for faster inference while applying the Mask RCNN model. Since, the inferring performance of the YOLOv3 is 3 times faster than the Mask RCNN on the actual image frame (with resolution 2304×1296) of the dataset. This shows a substantial evidence for the application of YOLOv3 model in implementing real time vehicle detection system. The experimental result is shown in Table 2.

Table 2: Comparison of inference speed for existing pre-trained model on different frame sizes

Model	Pretrained Architecture	Dataset	Inference time (ms)		
			720×576	2304×1296	3840×2160
Mask RCNN	Inception V2	COCO	12.77	26.31	57.47
YOLO V3	DarkNet-53	COCO	7.22	8.23	8.47

The Table 3 illustrates a comparison of the performance of two models YOLOv3 and Mask RCNN on three different visibility conditions i.e. foggy time, peak time and night time. Detection at Foggy time conditions and night time detection are considered as adverse environment detection in this work. The performance of each model is tabulated and then compared on the basis of mean average precision. From the table, it is noted that the overall mAP of YOLOv3 is better than that of Mask RCNN except in the night time when it is 10.26% than compared to 9.84% of YOLOv3. Likewise, the performance of YOLOv3 model for heavy vehicles i.e. truck and bus better than those of the Mask RCNN model. On the other hand, the performance of Mask RCNN model for the light vehicles i.e. motorcycle and car outclass YOLOv3 at all times with peak time as an exception where car of YOLOv3 betters Mask RCNN by a map number of only 1.26%. The highest mAP number as observed in the table is 52.38%, of YOLOv3 for truck during peak hours whereas the smallest is 0.09% of Mask RCNN's for truck during night. The data follows no particular pattern but more or less suggests that YOLOv3 model for heavy vehicles detection are better at all times and Mask RCNN model for light vehicles detection are better at all times with a very few exceptions.

Table 3: Comparison of $mAP_{0.5}$ between YOLOv3 and Mask RCNN

$mAP_{0.5}$	Foggy		Peak Time		Night	
	YOLO V3	Mask RCNN	YOLO V3	Mask RCNN	YOLO V3	Mask RCNN
Motorcycle	0.022	**0.0631**	0.0512	**0.0809**	0.0026	**0.0893**
Car	0.2711	**0.3745**	**0.373**	0.3604	0.2281	**0.2865**
Bus	**0.3978**	0.1083	**0.3447**	0.2324	**0.1405**	0.0335
Truck	**0.3136**	0.2783	**0.5238**	0.1045	**0.0226**	0.0009
Overall	**0.2512**	0.2061	**0.3232**	0.1945	0.0984	**0.1026**

YOLOv3 model top performs in peak time condition followed by performance in foggy and night time situation. While, Mask RCNN top performs in Foggy situation followed by the performance at peak time and night time situation.

We have further continued this comparison for the fittest available model. In mid 2020 YOLO provided its 4th and 5th version in a period of couple of months with further enhancements like cross stage partial network, spatial pyramid pooling, weighted residual connections, Drop-Block regularization, cross-iteration batch normalization's [28]. The enhanced YOLOv5 model is shown in figure 5. We used this latest update of YOLOv5 version (made on April 2021) in our comparisons before submitting the final paper in additional time given by the conference committee. The detection speed of YOLOv5 is not compared with previous models used since this experiment was performed on different machine.

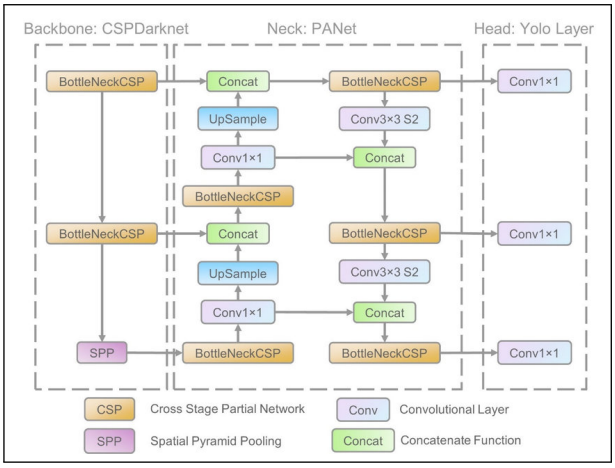

Figure 5: YOLOv5 [29]

Table 4: Comparison of $mAP_{0.5}$ between YOLOv3 and YOLOv5

$mAP_{0.5}$	Foggy		Peak Time		Night	
	YOLO v5	YOLO v3	YOLO v5	YOLO v3	YOLO v5	YOLO v3
Motorcycle	**0.08**	0.02	**0.22**	0.05	**0.19**	0.003
Car	**0.5**	0.27	**0.45**	0.37	**0.48**	0.23
Bus	**0.53**	0.4	**0.47**	0.34	0.12	**0.14**
Truck	**0.53**	0.31	**0.54**	0.52	**0.03**	0.02
Overall	**0.412**	0.25	**0.421**	0.32	**0.207**	0.09

The result shows that YOLOv5 top-performs its preceding model YOLOv3 in our dataset thus making it the fittest of all among the available state-of-art methods. This model is recent and is adopted to few vehicle detection tasks[30, 31]. YOLOv5 has 4 different types of model (namely small, medium, large, very large) on the basis of the complexity of the model. The model we used here for comparison is large model and was implemented using Pytorch[32]. The side by side accuracy comparison of YOLOv5 with the YOLOv3 is shown in table 4. YOLOv5 performed well in all conditions with all classes except for the Bus category during nighttime conditions which are highlighted in the table.

The overall mAP for both the models in night data is considerably very poor than in Peak hours and Foggy data. In addition, the performance of model is also degraded even for normal day light condition than on benchmark datasets. This could be because of characteristic of Heterogeneous Traffic moving in the same road in our collected dataset. Also, the detection accuracy for motorcycle class during all conditions is very poor using either of the models in our NTD.

5. Conclusion

In this work, we formulate a new custom dataset of Nepali Traffic called Nepali Traffic Dataset (NTD). Our data may be used in other experiments as an alternative data base or a difficult research collection, considering the sophistication of the proposed dataset.

Initially two types of comparisons: speed of model for detecting vehicle and its detection accuracy were made. YOLOv3 models outperforms Mask RCNN model with a little accuracy trade-off for the light vehicles detection in both type of comparison. However, YOLOv5 model outperform both the models in accuracy. From the above experimental result, we conclude that YOLO V5 architecture is a best choice among the other models for vehicle detection in our dataset.

Further, the Table 2 and 4 also illustrates that the vehicles detection accuracy of all the model is very low at night time than compared to other times which gives room for improving the accuracy for night time vehicle detection. Also, the performance of the models used for detecting motorcycle is very poor which identifies a huge research gap for improving the accuracy of the motorcycle detection.

As a future work, we will use YOLOv5 architecture for training, on our custom dataset, with some hyper-parameter tuning to design a novel model for detecting traffic vehicles in Nepal with better accuracy.

Acknowledgments

The authors express their sincere gratitude to Chorepatan Police Check Post and District Administration Office, Kaski for facilitating in

collection of surveillance video data. The authors also acknowledge the support of the faculty and administrative members from IOE-Paschimanchal Campus who have helped directly and indirectly to make this research successful.

References

[1] Tingting Huang. Traffic speed estimation from surveillance video data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[2] Ahmad Arib Alfarisy, Quan Chen, and Minyi Guo. Deep learning based classification for paddy pests & diseases recognition. In *Proceedings of 2018 International Conference on Mathematics and Artificial Intelligence*, pages 21–25, 2018.

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[6] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *arXiv preprint arXiv:1605.06409*, 2016.

[9] K He, G Gkioxari, P Dollár, and RB Girshick. Mask r-cnn. corr, abs/1703.06870. *arXiv preprint arXiv:1703.06870*, 2017.

[10] Tingyu Mao, Wei Zhang, Haoyu He, Yanjun Lin, Vinay Kale, Alexander Stein, and Zoran Kostic. Aic2018 report: Traffic surveillance research. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 85–92, 2018.

[11] Aleksandr Fedorov, Kseniia Nikolskaia, Sergey Ivanov, Vladimir Shepelev, and Alexey Minbaleev. Traffic flow estimation with data from a video surveillance camera. *Journal of Big Data*, 6(1):73, 2019.

[12] Yaoming Zhang, Xiaoli Song, Mengen Wang, Tian Guan, Jiawei Liu, Zhaojian Wang, Yajing Zhen, Dongsheng Zhang, and Xiaoyi Gu. Research on visual vehicle detection and tracking based on deep learning. In *IOP Conference Series: Materials Science and Engineering*, volume 892, page 012051. IOP Publishing, 2020.

[13] Fei Liu, Zhiyuan Zeng, and Rong Jiang. A video-based real-time adaptive vehicle-counting system for urban roads. *PloS one*, 12(11):e0186098, 2017.

[14] Guizhen Yu, Sifen Wang, Mingxing Li, Yaxin Guo, and Zhangyu Wang. Vision-based vehicle detection in foggy days by convolutional neural network. In *Chinese Intelligent Systems Conference*, pages 334–343. Springer, 2019.

[15] Md Nasim Khan and Mohamed M Ahmed. Trajectory-level fog detection based on in-vehicle video camera with tensorflow deep learning utilizing shrp2 naturalistic driving data. *Accident Analysis & Prevention*, 142:105521, 2020.

[16] Xuewen Chen, Huaqing Chen, and Huan Xu. Vehicle detection based on multifeature extraction and recognition adopting rbf neural network on adas system. *Complexity*, 2020, 2020.

[17] Tuan-Anh Pham and Myungsik Yoo. Nighttime vehicle detection and tracking with occlusion handling by pairing headlights and taillights. *Applied Sciences*, 10(11), 2020.

[18] D Sudha and J Priyadarshini. An intelligent multiple vehicle detection and tracking using modified vibe algorithm and deep learning algorithm. *Soft Computing*, 24:17417–17429, 2020.

[19] Qiao Meng, Huansheng Song, Yu'an Zhang, Xiangqing Zhang, Gang Li, and Yanni Yang. Video-based vehicle counting for expressway: A novel approach based on vehicle detection and correlation-matched tracking using image data from ptz cameras. *Mathematical Problems in Engineering*, 2020, 2020.

[20] M Hassaballah, Mourad A Kenk, Khan Muhammad, and Shervin Minaee. Vehicle detection and tracking in adverse weather using a deep learning framework. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[22] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 2020.

[23] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 2013.

[24] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov,

Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. *opencv/cvat: v1.1.0*, August 2020.

[25] Haojie Ma, Yalan Liu, Yuhuan Ren, and Jingxian Yu. Detection of collapsed buildings in post-earthquake remote sensing images based on the improved yolov3. *Remote Sensing*, 12(1):44, 2020.

[26] Sebastian Ramos Timo Rehfeld Markus Enzweiler, Rodrigo Benenson Uwe Franke Stefan Roth, Bernt Schiele Marius Cordts, Mohamed Omran, and B Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2016.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[28] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. *ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations*, April 2021.

[29] Renjie Xu, Haifeng Lin, Kangjie Lu, Lin Cao, and Yunfei Liu. A forest fire detection system based on ensemble learning. *Forests*, 12(2):217, 2021.

[30] Chengpeng Wang, Huanqin Wang, Fajun Yu, and Wangjin Xia. A high-precision fast smoky vehicle detection method based on improved yolov5 network. In *2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID)*, pages 255–259. IEEE, 2021.

[31] Margrit Kasper-Eulaers, Nico Hahn, Stian Berger, Tom Sebulonsen, Øystein Myrland, and Per Egil Kummervold. Detecting heavy goods vehicles in rest areas in winter conditions using yolov5. *Algorithms*, 14(4):114, 2021.

[32] Nikhil Ketkar. *Introduction to PyTorch*, pages 195–208. 10 2017.