The Performance Evaluations of Street Scene Detection by using Intelligent Object Detectors

Abbosjon Abdullayev[1] and Chang Gyoon Lim[2*]
Department of Computer Engineering,
Chonnam National University,
50 Daehakro, Yeosu, Jeonnam 59626, S. Korea
Tel. No.+82-61-659-7254, Fax No. +82-61-659-7259
abdullayev1705@gmail.com[1], cglim@jnu.ac.kr[2]
* corresponding author

Abstract- While Artificial Intelligence is being applied to security systems for security growth, the demand for valuable dataset is increasingly becoming huge. There are many open-source datasets for human and vehicle detection available on the Internet, however, there are limited datasets that include human body, human face, vehicle and Korean vehicle license plate as objects. To overcome this issue, in this work introduced a completely new well-annotated CNU Object Detection Dataset (CODD) to improve detection accuracy, especially in the Korean environment. CODD is a custom dataset with around 12,000 images composed of pedestrians and vehicles. There is a minimum of 2 objects and a maximum of 30 objects per image and about 84,000 object instances in the dataset. The validity and performance of our dataset were evaluated with various state of art architectures such as YOLO V3, YOLO V4, RetinaNet, and EfficientDet. The dataset demonstrates significant performance in the state of the art architectures. Thus, when the dataset employed to train models, significantly improved the detection accuracy.

Keywords – Surveillance System, Object Detection, Convolutional Neural Network, Data Analysis

I. Introduction

Object Detection is quite challenging and becoming an increasingly popular task in recent years. It is applied in broad areas of computer vision such as image searching, robotics, autonomous cars, Automatic License Plate Recognition (ALPR), Intelligent Transportation Systems, smart surveillance and security systems [1]. Almost all surveillance cameras are utilized to detect moving objects that appear in the video. In real-world content, humans and vehicles are the objects of interest for surveillance cameras and CCTV scenes. Hence, detecting humans and vehicles has attracted more attention in the research community. Numerous studies for understanding CCTV scenes showed that vehicles and humans are targeted as meaningful categories [2][3][4]. Based on this, persons can be recognized by their facial features and vehicle identification can be accounted for on the license plate.

The recognition of people and vehicles are essential in security and video surveillance systems. Therefore, it is pertinent to have quality datasets to meet up with the demand for high accuracy in training object detection models. However, there are limited open-source datasets that specifically include Korean vehicle license plate, human body, human face, and vehicle. To address the issue, we introduce a new complex dataset called CODD for the society of South Korea. CODD is a custom dataset with around 12,000 images composed of pedestrians and vehicles. There is a minimum of 2 objects and a maximum of 30 objects per image and about 84,000 object instances in the dataset. To understand the complex surveillance camera scenes, the objects are categorized into class.

The scenes were collected from real-life scenarios under different weather conditions, illuminations, shooting angles, background changing, etc., to ensure the detection of objects under various circumstances. Figure 1 shows the samples present in the dataset.

The main contributions of our study are summarized as follows:

- We introduce a new well-annotated object detection dataset named CODD. To the best of our knowledge, the dataset is the first publicly available one (with certain object categories) in the South Korean community.
- Experiments show that our CODD dataset yields high accuracy in real-world scene detection using state-of-the-art techniques such as YOLO V3[5], YOLO V4[6], RetinaNet [7], and EfficientDet [8].

Figure 1. Sample images from CODD dataset. In the images above, each target object is labelled with a 2D bounding box (BB)

The remainder of the paper is organized as follows: Section 2 presents the review of related works, section 3 discusses the CODD dataset in details, section 4 and 5 present experiments and the discussion of the result, while section 6 concludes the paper.

II. RELATED WORKS

Early developed methods concentrated on face detection using sundry special datasets. After the traditional face detection area [9][10] deep learning technics boosted the performance of face detection algorithms. Throughout the proposed algorithms, more clear and challengeable datasets like MegaFace[11], WIDER FACE[12], and benchmark FDDB[13] datasets were created. FDDB becomes a solution to the dataset volume and a wide range of difficulties such as low resolution, difficult poses, and poor lighting conditions. The dataset contains 2,845 images with a total of 5,171 labelled faces instances also grayscale images included.

In contrast, CODD produced 12,800 labelled faces instances and a certain part consists of faces with the mask.

In the field of video surveillance and automatic driving, Pedestrian detection is one of the important tasks. Several relevant datasets have been produced created, such as KITTI[14], Caltech-USA[15], in 2017[16]. The Caltech dataset is one of the big pedestrian dataset which consists of data recorded by a car BlackBox camera in an urban transportation environment. The database entails about ten hours of recorded video with a resolution of 640 × 480 pixels, 30 frames / second and around 250,000 2Dbounding boxes (including 2300 pedestrians and 350,000 rectangles). Furthermore, Zhang et al. presented a diverse pedestrian detection dataset named CityPersons in 2017[16]. Compared to some datasets, CityPersons has higher volume diversity and occlusion. In this work [17] to detect pedestrian in CCTV footage, the Faster R-CNN framework has been trained with the VOC2007[18] dataset. Another benchmark is CrowdHuman[19]. The dataset contains 470,000 human instances and each human instance is annotated with a full-body bounding box, the visible bounding box, and the head bounding box. In the data collection stage authors focused on dataset volume and crowd scenarios. For baseline detectors used Faster R-CNN[20] and RetinaNet which are based on the Feature Pyramid Network. Dataset showed better evaluation performance in crowd scenarios. LP detection is important in security and surveillance systems. One of the problems with LP detection is that its layout differs from one region to another. Efforts have been made to create datasets for many countries such as Indian, Turkish, American, Greek, Iraq, Israel, China and European datasets. LP dataset is normally described with several annotations such as LP bounding box, four vertices locations, LP number, and so on. Montazzolli et al. published the Automatic License Plate Recognition (ALPR) system based on CNN architectures(FAST-YOLO) in 2017 [21]. The model was trained using an SSIG-SegPlate dataset which contains 2,000 images with manually annotated 14,000 characters (alphanumeric symbols).

Other datasets in this category can be found in [22][23]. Some of the limitations in existing datasets include object distance from the camera, capture angle, backgrounds, and the lack of annotation. Our proposed CODD presents a new dataset with more sources and the number of instances per image.

III. CODD DATASET

In this section, we describe our CODD object detection dataset including the data collection process, annotation process, and instructive statistics. The process methodology is as described in Figure 2. The Crawling technique was employed for frame selection in the video. T is the initial time of the frame selection. The frames were extracted at an interval of 8 seconds

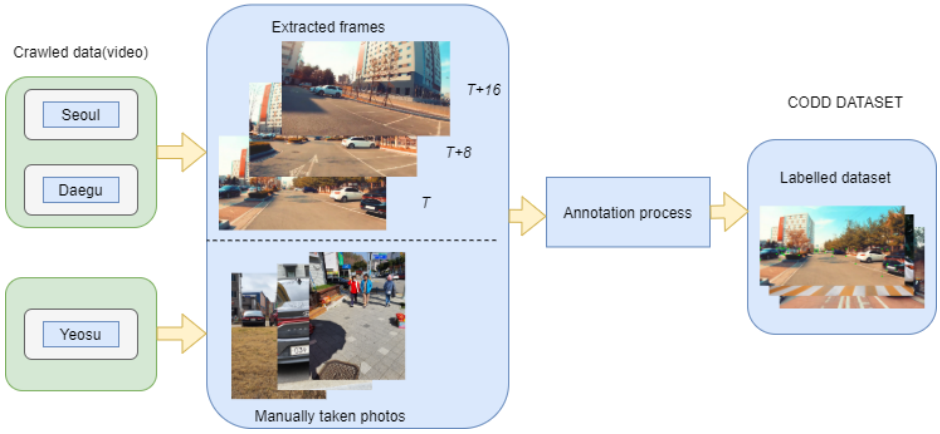

Figure 2. CODD dataset creation methodology T is the initial time where the selected frame

3.1 Data collection –

During the creation of our dataset, we paid attention to some criteria's which include: Image resolution, numerous viewpoints, backgrounds, different weather conditions, the density of objects in the image, different illuminations, full and cut objects. The range of the resolution of the images is from 600(width)×800(height) to1920(width)×1080(height). Different angles were used to gives us video footage from various viewpoints which will help models to detect objects from any viewpoint. The background considered includes tunnels, highways, streets, and beaches. The Crowded scenes with various degrees of object density such as marketplace, nightlife zones and office districts were considered. The images were capture in cloudy, sunny, rainy, snowy, day and night to give a variety of illumination in the dataset. We also annotated full and cut objects to enable models to detect objects even if they are cut in the frame or partially visible.

The dataset was collected in 3 different cities in Korea in which two of those, are known for traffic jams. About 4,000 images were collected by a handheld camera in Yeosu city. The advantage of a handheld camera is that it allows you to take pictures at any angle one wants. All the pictures collected in Yeosu city were carefully collected in one month, taking into account the rainy, cloudy, sunny conditions of the day and night between 7 am and 9 pm. Furthermore, the points we have considered are partially mentioned in a few studies [21][22][23]. The objects are occluded in the range from 30 to 100 percent in those images.

We collected videos, which have been recorded via a built-in black-box camera installed in the vehicles in big cities of South Korea such as Seoul. Moreover, we crawled videos that have been recorded in vibrant streets from Daegu city by walking, bicycling with the installed camera on the body. Overall, 68 videos were crawled. The crawling technique increases the diversity of our dataset. Thus, the difference time between extracted frames is about 8 seconds (per extracted frame to per 8 seconds) or 250 frames (per extracted frame to per 250 frames). After the process, 7,000 unique frames were collected. All data were taken in a moving vehicle which indicates that the data is more applicable in real life conditions. Additional 1,000 photos were selected from publicly available data from well-known car dealerships in Korea.

3.2 Image annotation

Labelling is the most time-consuming task in the data pre-processing pipeline. The images were manually annotated using the open-source graphical annotation tool LabelImg [24]. The tool was used to label the visible target objects in the dataset with 2D bounding boxes. This means that every image is annotated by a rectangle with

(x_c, y_c, w, h) coordinates where x_c, y_c is the centre of the boxes w and h rectangle width and height, respectively. The manually labelled data were then saved as XML files in PASCAL VOC format as well as in YOLO format. For this study, four objects were label from 0 to 3 as shown in Table 1.

Table -1: Categorization of objects

Class_id	0	1	2	3
Class_name	Car	LP	Person	Face

Figure 3 shows a sample annotated frame with about 13 object instances. In most cases, our frames consist of more than 10 object instances which is consistent with real-life scenes.

Figure 3. Annotation process. (a) The top left green point represents the starting point of every bounding box annotation. (b) Total 13 object instances labelled, and three type categories have occurred

3.3 Dataset statistics-

The CODD consist of 12,043 unique annotated images. Due to the variety of sources of our data, the images have various resolutions. Most of the data resolution is 1,920(Width) x 1,080(Height) x 3(Channels). A total of 84,000 instances were manually labelled in the dataset with each category having enough representation as shown in Figure. 4. Thus, CODD is well-balanced in terms of instances for each category. Figure 5 presents the comparison of our dataset with other popular available datasets in terms of instances and categories per image. On average, CODD consists of 1.75 categories and 7 instances per image, while PASCAL VOC consists of 1.5 categories and 3 instances

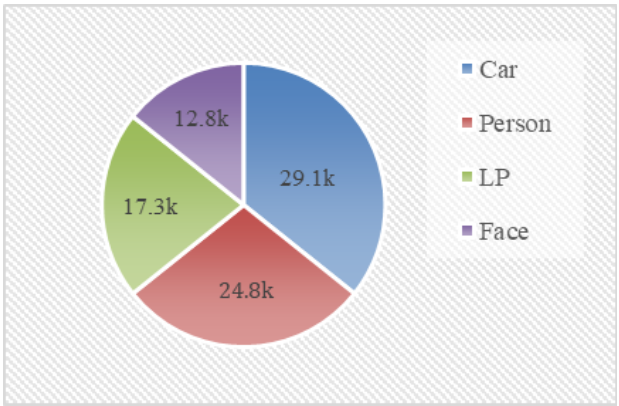

Figure 4. Numbers of instances per category

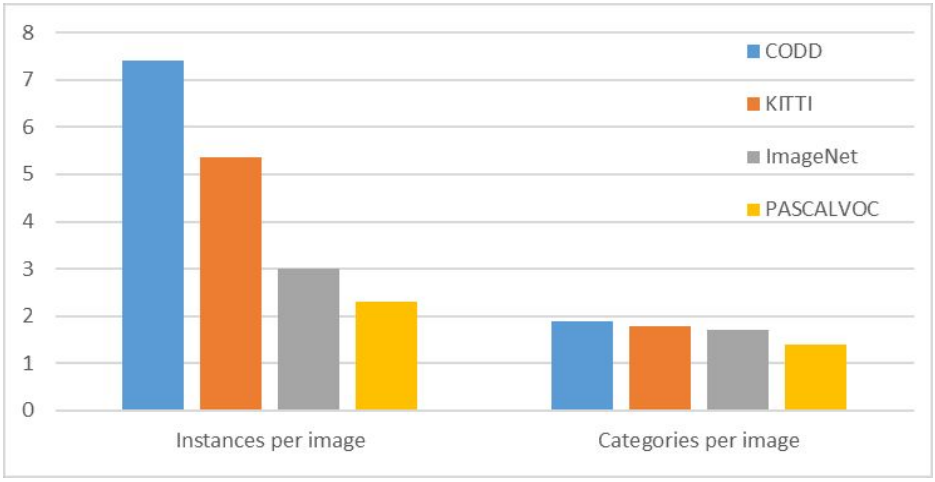

Figure 5.Comparison of statistical results of CODD and other public datasets in term of instances and categories per image

Moreso, only about 9 % of CODD images consist of only one category- This is important as training the model with data that has more categories will give a better performance. Considering the LP, CODD has the highest label instance of 12,800, while KARPLATE, NI-VI, GTI datasets contain 6000, 1500 and 3425 respectively.
Likewise, CODD has the highest figures in the instance/image, cities and seasons of 1.53, 4 and 3 respectively, as indicated in Table 2. Furthermore, the CODD occlusion level was made well balance by annotating any objects with 30% or more visibility within a frame.

Table -2. Comparison CODD with public datasets by LP component

	GTI(ESP)	NI-VI(Iraq)	KARPLATE(Kr)	CODD(Kr)
num.of labeled instances	3,425	1,500	6,000	12,800
num.of instances/image	1	1	1.5	1.53
num.of various of cities	1	2	2	4
num. of various of seasons	1	1	1	3

Table 3 presents the comparison of CODD with other datasets in terms of category. As indicated, our CODD includes all four categories. Meanwhile KITTI, PASCALVOC and MSCOCO[25] consist of only two (person, car), while CITYSPACES and KARPLATE consist of three (Person, Car, LP) and one (LP), respectively. Besides, faces category instances included faces with masks and those without masks. Faces with masks amount to about 4,680 instances while those without masks are around 8,100 instances.

Table -3. CODD object categories in other publicly available datasets

Categories	MSCOCO	PASCALVOC	KITTI	KARPLATE	CITYSPACES	CODD
Person	✔	✔	✔	×	✔	✔
Faces	×	×	×	×	×	✔
Car	✔	✔	✔	×	✔	✔
License Plate	×	×	×	✔	✔	✔

IV. EXPERIMENTS

In this section, we was utilized the CODD dataset to train state-of-the-art one stage detectors and we compared the results with each other by. All processes in this section are shown in Figure 6. Experiments were conducted on a virtual machine with GPU Tesla V100-SXM2-16GB on the Google Colab platform

4.1 Selected models for the experimeints-

To understand the quality of the created CODD dataset, state-of-art classifiers such as RetinaNet, YOLO V3, YOLO V4, and EfficientDet were trained and tested. The 12,043 images of the CODD were randomly split into 9,200, 1,500 and 1,343 unique images for training, validation and testing the model, respectively.

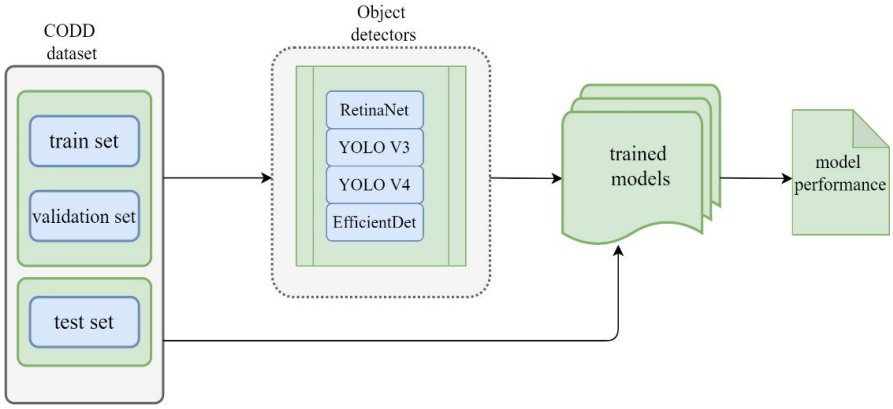

Figure. 6. The Selected models experimental processes on CODD dataset

The models were built for the multiclass classification problem. The multiclass object for the detection includes faces, person, car and license plate. All experiments used Python 3.6 and were based on the Tensorflow framework.

4.2 Training models-

We first trained RetinaNet dense object detector on CODD where the model addresses the extreme foreground-background class imbalance issue by modifying the loss. Architecture adopted ResNet-50[26] backbone with a Feature Pyramid Network (FPN)[27]. As already mentioned, the model builds in the virtual machine Google Colab. After data preprocessing, The model was trained on 200 epochs with batch size and steps of 8 and 7,000, respectively. End of each epoch pre-trained weights saved to snapshots directory where located in google drive.
Next, we trained with YOLO V3 and V4 models. As the backbone, for both YOLO versions used Darknet[28]. The input shape image resolution is resized to 416x416 pixels for optimal performance on our machine. The training was conducted using batch size and subdivision of 64 and 16 respectively. Max batches set 8,000 and steps=6400, 7200 values were given respectively. To extract more high-level futures, we use 27 filters before the yolo layers. Normally in Darknet platform number of filters calculated by filters = (number of classes+5) * 3 by this this equation.
Finally, we produced 9,200 images and their annotations to the network for the training. It took around16 hours to train the model with 9,000 epochs Every 1,000 epochs pertained weights saved to Google drive to keep track of accuracy and losses.
Recently the google brain team produce a new model named EfficientDet which the authors focused on model efficiency by proposing a weighted bi-directional feature pyramid network (BiFPN)[8] and a new Compoundscalingmethod. Lastly, the EfficientDet-D1 version of EfficientDet on Efficientnet-B1 backbone was employed was trained using SGD optimizer with weight decay and momentum of 4e-5 and 0.9, respectively. The Learning rate changed during the training in the range [0,0.16]. furthermore, due to the version of the model, the image input size was set to 640x640 and trained on 300 epochs. The model training took about 10 hours to complete. Table 4 shows the summary of the selected training key hyper parameters and training hours with number of epochs for four object detectors.

Table -4. Hyper-parameters used in training process

Hyper-Parameters	RetinaNet (ResNet-50)	YOLOV3 (Darknet-53)	EfficientDet (BiFPN)	YOLOV4 (Darknet)
Initial learning rate	0.1	0.001	0.16	0.01
Batch size	16	64	32	64
Momentum	0.9	0.949	0.9	0.949
IOU threshold	0.5	0.6	0.5	0.6
Epochs	200	9000	300	9000
Training hours	8 hours	16 hours	10 hours	16 hours

4.3 Metrics for the formance evalution -

We randomly selected 1343 images from the dataset as the test set. The test set contains 3320 car instances, 2246 person instances, 1789 LP, and 794 face instances. The trained models were all tested on this same set of test data.

We used mean average precision (mAP) and average precision (AP) to evaluate the performance of the object detection algorithms. Also, we computed evaluation metrics like Recall, F1 score, Precision [29] to ensure more quality of detectors. In Yolo versions, the threshold from PASCAL VOC is set to 0.6 as IOU. As compared to previous works [1, 14, 16, 20] that set IOU to 0.5, we set 0.6 to detect with more clarity in the YOLO V3, V4.

V. RESULTS AND DISCUSSION

Figure 7 shows the detection results of each model. The models were able to categories the target objects on rainy road scenes as indicated in Figures (a),(b), (c) and (d). Detection results show each visible target object has been detected in all four models. However,one of the LP showed lower detection accuracy, especially, that is observed on RetinaNet baseline. This is since the image was taken in rainy conditions.

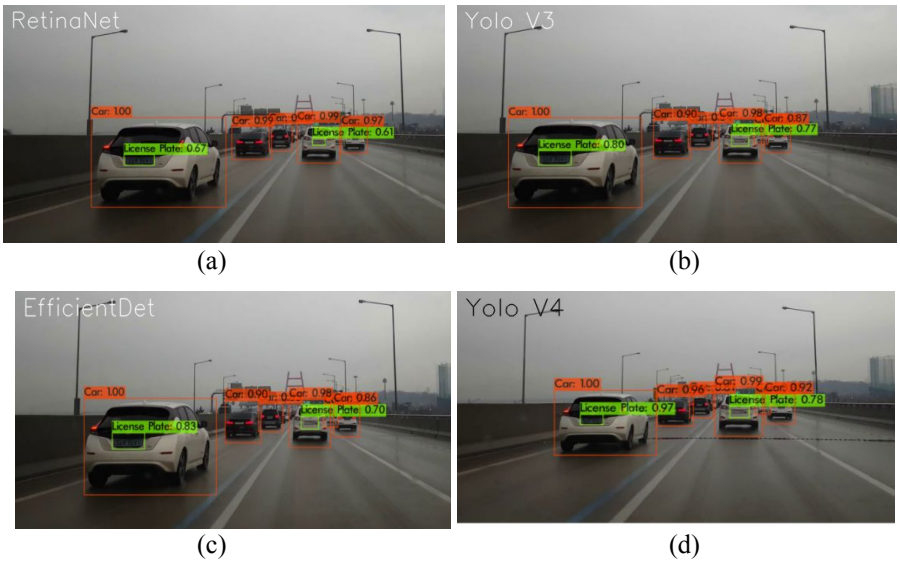

Figure. 7. The detection results for the four state-of-art algorithms on highway footage. Green BB indicates LP with detection rate and orange BB shows car category with accuracy The texts in the upper left corner of the images indicate the name of the architecture

Figure 8 shows the performance of the models on street scene data. The models' were able to categorize Persons with mask and without a mask with a good performance. Also, the models' correctly categorize people from the backside**.**

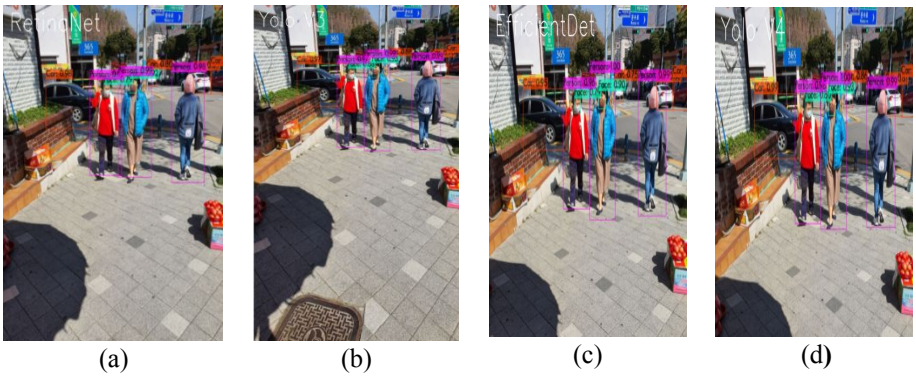

Figure 8 The detection results of the four state of the art algorithms in street footage

Results in Figure 9 (a, b, c) show the CODD dataset properly balanced with the diversity of backgrounds and instances of objects in various positions In Figure 9 (a) car detected the noteworthy ground truth BB is drawn with high IOU even when the car door is open. Figure 9 (b, c) also reached excellent detection results accordingly (snowing and night conditions). In Korea, various types of LP. In Figure 10(a, b) detected two types of LP's and encountered all cars. Besides, the CODD dataset supports all common types of LPs

(a) (b) (c)

Figure 9. The detection results of the four state of the art algorithms in car parks.In (a, b, c) under real life conditions:In the images some part of LPs were blurred to protect privacy

In Figure 10 (b) it counted 6 cars though at first glance 4 cars are visible, the remaining 2 automobiles are also very small and partially visible.

(a) (b)

Figure 10. The detection results under following conditions: Various shooting distances, multiple objects, various types of License Plates. The texts in the upper left corner of the images indicate the name of the architecture. In the images part of the fully visible LPs and Faces were blurred to protect privacy

5.1 Performance evaluation-

The performance levels of the selected algorithms for each category shown in Figure 11, it can be seen that the detection of the cars has an advantage among all the categories.The car category achieves better performance in the

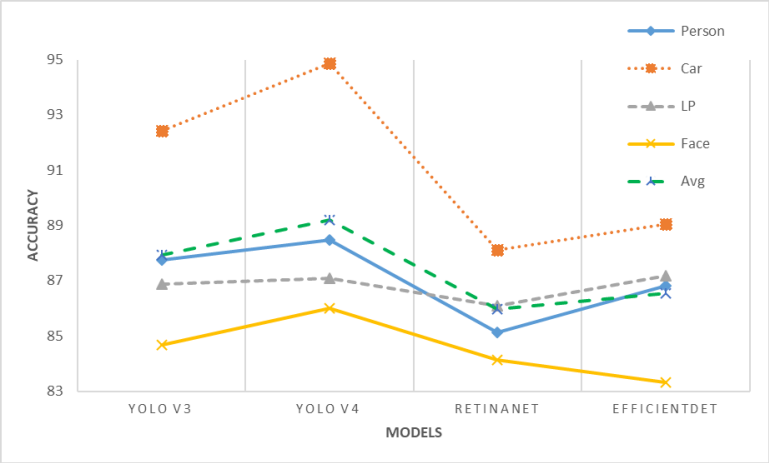

Figure 11. Performance evaluation of the four state –of –art models for each category

YOLO V4 model, in RetinaNet architecture gets worse performance (≈ 88 percentage). To analyze the Person category YOLO V4 has also a larger advantage than other models and lower performance shown on the RetinaNet model. However, in the LP category, EfficientDet gets better performance than other models displaying a 1.10 percentage higher performance than RetinaNet. FPN based RetinaNet produces weaker performance in Face detection scenarios with 83.33 percentage. To sum up processes, YOLO V4 serves as a better feature extraction detector in most categories. Figure 12 shows the more Intersection Over Union all categories increase the more The AP value decreases. When IOU=0.9 person is, LP and face categories get the worst results. In the car category, it detects 33 percent of clarity with IOU=0.9 adjustment..

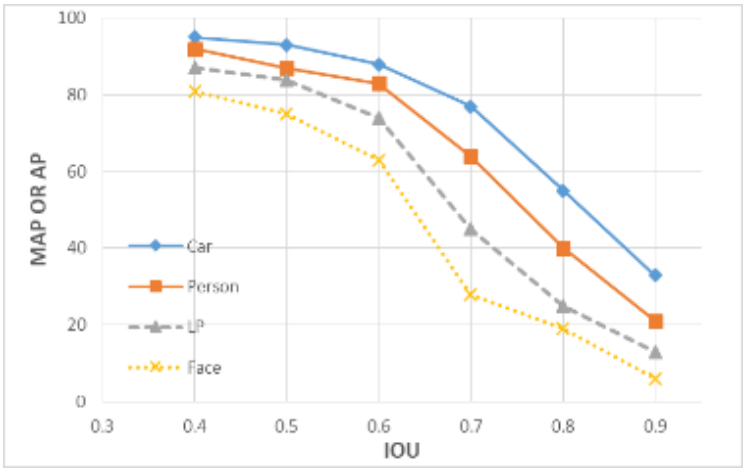

Figure 12. Performance evaluation of the YOLO V4 baseline model for each category

The performance of the models for each category is shown in Table 5. The performance of the models ranges between. 83.33 to 94.87 detection rate. From the results in Fig 10, Car was the best-detected object by all the models with RetinaNet and YOLO V4 having the least and best performance, respectively. Besides, the face was the least detected object by all the models with RetinaNet and YOLO V4 having the least and best performance, respectively. LP and Person maintain the detection performances of the range 86.09-87.19 and 83.33-86.01 across the models,

respectively. Consequently, YOLO V4 offers better performance in the detection of Car, Person and Face. Meanwhile, EfficientDet offers the best performance for analyzing LP.

Table -5 (AP) detection rate for every category by appropriate

	Yolo V3	Yolo V4	RetinaNet	EfficientDet
Person	87.75	88.46	85.14	86.8
Car	**92.43**	**94.87**	88.1	89.04
LP	86.89	87.09	86.09	87.19
Face	84.68	86.01	84.13	83.33
Avg	**87.93**	**89.2**	**85.97**	**86.54**

5.2 The Models Average Detection Performance Evaluation -

The overall detection accuracy of the models was evaluated using the recall, precision, F1 score, IOU, mAP and accuracy as presented in Table 6. In terms of recall, YOLO V4 and RetinaNet performed the best and least with a score of 88.4 and 84.7, respectively. Similarly, YOLO V4 has the highest precision of 89.77 while RetinaNet has the least precision of 86.09. Consequently, YOLO V4 recorded the overall best performance of 89.08, 75.22, 89.34 and 89.2 in terms of F1 score, IOU, mAP and accuracy, respectively, as presented in Figure 13. The Least performance was recorded by RetinaNet across all the performance indicators.

Table -6. Models average performances with CODD

Method	Recall	Precesion	F1 score	Avg IOU	mAP	Avg Acc
RetinaNet	84.7	86.13	85.4	63.3	86.89	86.54
YOLO V3	87.5	87.2	87.34	73.63	88.14	87.93
YOLO V4	88.4	89.77	89.08	75.22	89.34	89.2
EfficientDet	85.3	86.09	86.77	70.2	87.04	85.97

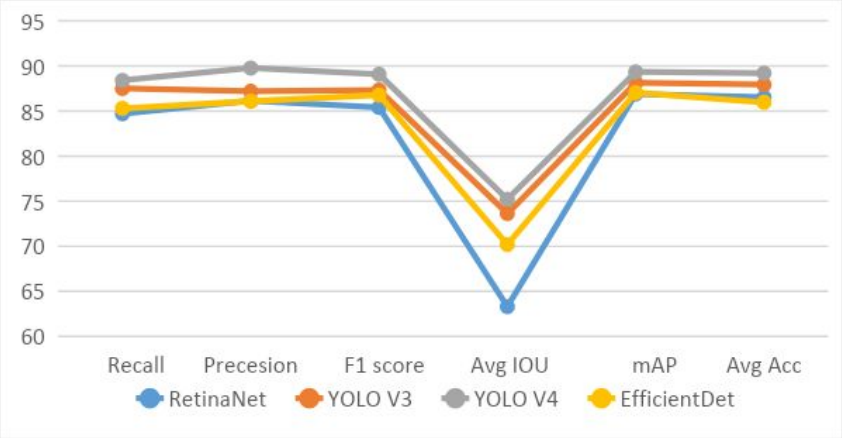

Figure 13. Graphical visualization of metric results

VI.CONCLUSION

In this work, we introduce a new object detection dataset named CODD. The dataset contains 4 types of object categories such as a car, person, face, and license plate in Korea. Images were collected in various cities and on different days and includes uncontrolled conditions. Each image is carefully annotated and finally, around 84,000 object instances are labeled. Statistical analysis show CODD achieve better results in term of density and diversity among the compared datasets. State-of-the-art technics which trained on the CODD dataset give better results in understanding street scenes, especially crowded scenarios. However, there is still room for improvement in terms of the number of categories as we only considered four in this work. Categories like Expanding the dataset volume by adding new images continuously, balancing all selected class instances and expansion of camera positions can be considered in future works. The accuracy of detection can also be improved through careful design and tuning of the models

REFERENCES

[1] Abdul Vahab, Maruti S Naik, Prasanna G Raikar "Applications of Object Detection System," in International Research Journal of Engineering and Technology (IRJET).e-ISSN: 2395-0056 ,2019

[2] A. Dimou and F. Alvarez, "MULTI-TARGET DETECTION IN CCTV FOOTAGE FOR TRACKING APPLICATIONS USING DEEP LEARNING TECHNIQUES ı a Universidad Polit ´ Tecnica de Madrid (GATV), Spain Information Technologies Institute, Centre for Research and Technology Hellas, Greece," pp. 3–7.

[3] A. P. Shah, J. B. Lamare, T. Nguyen-Anh, and A. Hauptmann, "CADP: A Novel Dataset for CCTV Traffic Camera-based Accident Analysis," *Proc. AVSS 2018 - 2018 15th IEEE Int. Conf. Adv. Video Signal-Based Surveill.*, no. I, pp. 1–9, 2019, DOI: 10.1109/AVSS.2018.8639160.

[4] K. B. Lee and H. S. Shin, "An Application of a Deep Learning Algorithm for Automatic Detection of Unexpected Accidents under Bad CCTV Monitoring Conditions in Tunnels," *Proc. - 2019 Int. Conf. Deep Learn. Mach. Learn. Emerg. Appl. Deep. 2019*, pp. 7– 11, 2019, DOI: 10.1109/Deep-ML.2019.00010.

[5] J. Redmon and A Farhadi "YOLOv3: An Incremental Improvement." 2018, *arXiv*:1804.02767.[Online]. Available : http://arxiv.org/abs/1804.02767

[6] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection,". 2020. *arXiv:*2004.10934.[Online] Available: https://arxiv.org/abs/2004.10934

[7] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020, DOI: 10.1109/TPAMI.2018.2858826.

[8] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 10778–10787, 2020, DOI: 10.1109/CVPR42600.2020.01079.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, HI, USA, 2001, pp. I-I, doi: 1 0.1109/CVPR.2001.990517. [10] M. Yang, D. J. Kriegman, S. Member, and N. Ahuja, "Detecting Faces in Images : A Survey," vol. 2 4, no. 1, pp. 34–58, 2002.

[10] Ming-Hsuan Yang, D. J. Kriegman and N. Ahuja, "Detecting faces in images: a survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 34-58, Jan. 2002, doi: 10.1109/34.982883.

[11] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller and E. Brossard, "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 4873-4882, doi: 10.1109/CVPR.2016.527

[12] S. Yang, P. Luo, C. C. Loy and X. Tang, "WIDER FACE: A Face Detection Benchmark," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 5525-5533, doi: 10.1109/CVPR.2016.596.

[13] V. Jain and E. G. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings", 2010."

[14] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics : The KITTI Dataset," vol. 32, no. October 2011, pp. 1–6, 2013.

[15] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, 2012, DOI: 10.1109/TPAMI.2011.155.

[16] S. Zhang, R. Benenson and B. Schiele, "CityPersons: A Diverse Dataset for Pedestrian Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 4457-4465, doi: 10.1109/CVPR.2017.474

[17] A. Dimou, P. Medentzidou, F. Á. García and P. Daras, "Multi-target detection in CCTV footage for tracking applications using deep learning techniques," 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, 2016, pp. 928-932, doi: 10.1109/ICIP.2016.7532493.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010, DOI: 10.1007/s11263-009-0275-4.

[19] S. Shao, Z. Zhao, B. Li, T. Xiao, and G. Yu, "CrowdHuman: A Benchmark for Detecting Human in a Crowd,".2018. pp. 1–9.

[20] R. Gavrilescu, C. Zet, C. Foşalău, M. Skoczylas and D. Cotovanu, "Faster R-CNN:an Approach to Real-Time Object Detection," *2018 International Conference and Exposition on Electrical And Power Engineering (EPE)*, Iasi, 2018, pp. 0165-0168, doi: 10.1109/ICEPE.2018.8559776.

[21] P. Alegre, "License Plate Detection and Recognition in Unconstrained Scenarios."2018.

[22] S. Jabri, M. Saidallah, A. El Belrhiti El Alaoui, and A. El Fergougui, "Moving Vehicle Detection Using Haar-like, LBP and a Machine Learning Adaboost Algorithm," in *IEEE 3rd International Conference on Image Processing, Applications and Systems, IPAS 2018*, 2018, no. March, pp. 121–124, DOI:.10.1109/IPAS.2018.8708898.

[23] N. O. Yaseen, S. G. S. Al-Ali, and A. Singer, "Development of New AN Dataset for Automatic Number Plate Detection and Recognition in North of Iraq," *1st Int. Informatics Softw. Eng. Conf. Innov. Technol. Digit. Transform. IISEC 2019 - Proc.*, pp. 4–9, 2019, DOI: 10.1109/UBMYK48245.2019.8965512.
[24] *LabelImg*,Accessed: Nov.1, 2019.[Online].Available: https:https://github.com/tzutalin/labelImg
[25] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," *Lect. Notes Comput. Sci. (including Subsea. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014, DOI: 10.1007/978-3-319-10602-1_48.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "ResNet," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016.

[27] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2017, DOI: 10.1109/CVPR.2017.106.

[28] R. Lewenstein and R. Lewenstein, "Darknet," in *Darknet*, 2018.

[29] P. A. Flach and M. Kull, "Precision-Recall-Gain curves: PR analysis is done right," 2015.