2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

Real-time Pedestrian and Vehicle Detection for Autonomous Driving

Zhiheng Yang, Jun Li, and Huiyun Li, *Member, IEEE*

Abstract— **Fast and efficient pedestrian detection and vehicle detection has become an increasingly important task in the autonomous driving technology. In this paper, we propose a new pedestrian detection and vehicle detection algorithm based on the YOLOv2 with optimized feature extraction. We adopt the priori experience about the feature box sizes, instead of K-mean clustering algorithm in the original YOLOv2 algorithm. We first conduct statistical analysis on the dataset with a label of pedestrian label and vehicles, and then we design the initial value of the pre-selection box that is more in line with the characteristics of pedestrian and vehicle. Together with hard negative mining, multi-scale training, and model pretraining, the proposed algorithm not only improves the detection accuracy but also keeps the good detection efficiency. Experimental results on traffic benchmark record demonstrate that the optimized algorithm satisfies the real-time capability and the accuracy requirement of the low-speed autonomous driving.**

Keywords- pedestrian detection, vehicle detection, YOLOv2, real-time, Darknet-19

I. INTRODUCTION

With the rapid development of autonomous driving technologies, fast and efficient detection algorithm becomes an increasingly important task [1]. Compared with ultrasonic, laser, radar, and other technologies, visual detection is more in line with the habit of human eyes to capture information and is much less expensive to implement. Vision-based object detection technology uses the vehicle sensor to sense the surrounding environment of the vehicle and controls the steering and speed of the vehicle. According to the information of the road, the position and the obstacle acquired by the sensor to provide the vehicle accessible area, vehicles can drive on the road safely and reliably.

The target of intelligent driving in vehicle vision includes vehicles, pedestrians, traffic signs, etc. Among them, pedestrians and vehicles are two of the main participants in the road traffic environment. Pedestrians are not rigid objects; they are variable in posture and appearance. Due to the problems of the posture, distance, and occlusion, the vehicle cannot be detected accurately, sometimes. The background environment is also versatile, increasing the difficulties of recognition and detection. In particular, the real-time on-board pedestrian and vehicle detection system is an important part of the automatic assisted driving system. An accurate and real-time detection system with small volume and low power consumption can effectively protect the pedestrians and vehicles.

There were numerous research efforts about accurate pedestrian detection and vehicle detection in the field of traditional vision. Traditional object detection generally uses a sliding window framework, and most models of them employed image processing to extract the visual features associated with the candidate regions, such as Viola-Jones (VJ) [2] and Histogram of Oriented Gradient (HOG) [3]. These models calculated gradient-direction histogram features or Haar-like features of local regions as classifier inputs but were easily affected by light and occlusion. Then, the Deformable Part Based Model (DPM) [4], which is developed from the HOG detector, can effectively solve the occlusion problem. It used a combination of multiple correlation sub-models and improved the performance compared to previous methods significantly, but the effect is always unsatisfactory.

The success of ImageNet challenge [5] and Pascal visual object class (VOC) challenge [6] showed that object detection based on deep learning has good development prospects. For example, the Region-based Convolutional Neural Networks(R-CNN) [7] is an object detection method combined with Region Proposal and Convolutional Neural Network (CNN). Compared to the previous best results, it has improved the Mean Average Precision (mAP) by more than 30%. The object detection methods related to deep learning can also be roughly divided into two groups: 1. Region Proposal such as R-CNN, SPP-net [8], Fast R-CNN [9], Faster R-CNN [10], R-FCN (Region-based Fully Convolutional Networks) [11]; 2. End-to-End (no Region Proposal), such as YOLO (You Only Look Once) [12], SSD (Single Shot multi-box Detector) [13]. The kinds of deep learning based on Region Proposal, typically obtain a Region of Interest (ROI) by using Selective Search(SS) or Region Proposal Network(RPN) and then classify each region using the CNNs model to get the category and confidence level. Another ones adopt a single-channel network architecture to combine the object box with object identification, and output the target category and the target location information at one time, those algorithms enhance the execution efficiency of the identification. At present, the methods based on the Region Proposal still prevail, but the advantages in the detection speed of the end-to-end method

Zhiheng Yang is with the School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin, 541004, China. (e-mail: zh.yang@siat.ac.cn).
Jun Li is with the Guangxi Key Laboratory of Optoelectronic Information Processing, Guilin University of Electronic Technology, Guilin, 541004, China (e-mail: lijun09@ guet .edu.cn) .
Huiyun Li, Shenzhen Institutes of Advanced Technology, The Chinese University of Hong Kong, Chinese Academy of Science, Shenzhen 518055, China (corresponding author to provide e-mail: hy.li@siat.ac.cn).

978-1-5386-4451-5/18/$31.00 ©2018 IEEE

are obvious and the follow-up development will wait to be seen.

This paper proposed a detection system based on the YOLOv2 [14], and it gets pedestrian-friendly anchor boxes easily by adopting priori experience about the feature box sizes. The system combines several important strategies, including hard negative mining, and multi-scale training, et al. to improve the detection accuracy. Then, the KITTI Object Dataset [16] uses to tone the ConvNets parameters with pedestrian and vehicle features. Compared with the Faster R-CNN and YOLOv2, proposed system achieve better results in detection accuracy and is as faster as the YOLOv2, so it is enough for real-time applications in the autonomous vehicle driving.

The rest of this paper organized as follows. Section 2 provides a review of the related work on pedestrian detection and vehicle detection. Then the proposed detection system described in Section 3. Experimental results are given in Section 4. Section 5 provides a summary of the paper and future work.

II.RELATED WORK

Pedestrian detection and vehicle detection is an important branch of object detection in the field of automatic driving applications. The existing methods are divided into two types of traditional model and deep learning model. Traditional detection generally uses the sliding window framework, which mainly includes three steps:

- Step1：Using different size of the sliding window to choose a part of the map as a candidate area;
- Step2：Extract the visual features associated with the candidate area. Such as Harr features commonly used in face detection; HOG features commonly used in pedestrian detection and ordinary target detection;
- Step3：Use classifier to identify, such as the common Support Vector Machine (SVM) model.

Dalal et al. proposed gradient direction histograms (HOG) descriptors to extract contour features. Viola and Jones extract two Haar-like features based on Haar-like features and extensions. Viola-Jones (VJ) and Histogram of Oriented Gradient (HOG) are handle models, and they provide reasonable results with the MIT pedestrian database [15] and the INRIA database [3]. However, these methods are susceptible to both light and target occlusion and perform poorly with newer and more challenging datasets such as the KITTI Object Dataset the Caltech Pedestrian Dataset [17]. In traditional object detection, the DPM, which significantly improves the performance, compared to previous methods and won consecutive VOC 2007-2009 detection champion. The DPM treats objects as multiple components and describes the objects in terms of the relationships between components, this feature that best suits the non-rigid nature of many objects in nature. The DPM has achieved good results in some detection tasks, but the DPM is relatively complex, the detection speed is slower.

Recently, the deep-learning model has improved the performance of object detection; especially the detection model based on the Region Proposal provides a good object recognition ability. Ross Girshick et al. combined the CNNs with the Region Proposal. They chose Selective Search to extract the region proposals, adopt the CNNs to extract features and used an SVM as the classifier. It applied CNNs to translate detection problems into classification problems that improved the mAP by more than 30% relative, but this method has the disadvantages of repeated calculation, complicated training. He Kaiming et al. proposed SPP-net, added a space pyramid pool layer (SPP layer) before the full-connected layer to lift the fixed size constraint. The extracted features have better scale invariance and reduce the possibility of overfitting. Then, Ross Girshick et al. designed a simple and scalable detection algorithm called Fast R-CNN, and this method only uses the CNN for feature extraction, and uses a special pooling layer, like an SPP layer, and two full connection layers instead of SVM to achieve classification, which greatly reduces the computational complexity and training complexity. Although these methods have greatly improved the recognition accuracy, they still need to be improved in terms of implementation efficiency. In order to combine preselected box extraction and object classification into a network framework, Ross Girshick et al. designed the RPN, which used a pre-trained CNNs to generate propose regions form the image, quicker than the selective search or edge boxes. Then this method combined with the Fast R-CNN, called Faster R-CNN, further improved the implementation efficiency, and meet the real-time requirements. The R-FCN changes the last fully connected layer to a position-sensitive convolutional network so that all calculations are shared. Of cause, other End-to-End algorithms have a high efficiency of implementation, such as YOLO, SSD and so on. The YOLO adopted a single-channel network architecture to combine the object box with the identification, and outputted the target category and the target location information at one time, which enhances the execution efficiency of the identification. However, the recognition accuracy to be improved, especially for the identification of small targets. The SSD also uses a single CNN network infrastructure and learn the Anchor mechanism from Faster R-CNN using multi-size feature-pyramid prediction, which identified in different levels of the feature maps. This enhances the recognition of small objects. The YOLOv2 uses a series of methods, such as multi-scale training, fine-grained features, and batch normalization, convolutional with anchor boxes, dimension clusters, and et.al, to improve the original YOLO multi-object detection framework. Then the basic model of YOLOv2 uses a framework called Darknet-19 as a feature extractor, based on previous work experience. The Darknet has 19 convolutional layers and 5 max-pooling layers, which is a similar GoogleNet network structure. In addition, the Darknet adds the pass-through layer, similar to ResNet [18] combining high-resolution features with low-resolution features, to add finer grained features. The YOLOv2 uses K-means clustering to select automatically the best initial boxes. To increase the intersection over union (IOU) score of the selected boxes, the YOLOv2 define the equation (1):

$$d(box, centroid) = 1 - IOU(box, centroid) \quad (1)$$

According to the standard k-means with Euclidean distance result, select the number of initial anchor boxes, k = 5. Those measures can improve the precision with the advantage of keeping the original speed. Figure 1 provides an approximate comparison of the performance of various methods (frames per second, FPS) and versus MAP, according to relevant data collected from the different papers. Due to the difference between the evaluation hardware and the environment, the data is for reference only and not an absolute contrast.

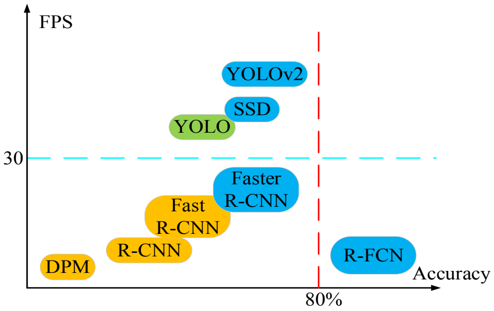

Figure 1. The comparison of different methods in accuracy and efficiency.

III. THE PEDESTRIAN DETECTION SYSTEM

A. *Proposed Model*

The proposed detection system in this paper has the similar architecture of the YOLOv2, which has been proved to be one of the state-of-the-art deep learning schemes for generic object detection. The proposed networks adopt a single CNN channel network structure and the suitable anchor-boxes convolution, to predict the object's location and classification directly.

In this work, we propose to extract the anchor boxes architecture for pedestrian detection and vehicle detection. In view of the characteristics of the detection system, we adopt the priori experience about the sizes of the feature boxes, instead of K-mean clustering algorithm in the proposed model to reduce the computational complexity. We conduct statistical analysis the pedestrian and the vehicles' labels on the dataset, and design the initial value of the anchor boxes that is more in line with the characteristics of pedestrian and vehicle, as shown in Figure 2. And we train and test the proposed model, combining a number of strategies to improve pedestrian detection accuracy, including hard negative mining, multi-scale training, model pretraining, and proper calibration of key parameters. The proposed detection model training procedure is shown in Figure 3.

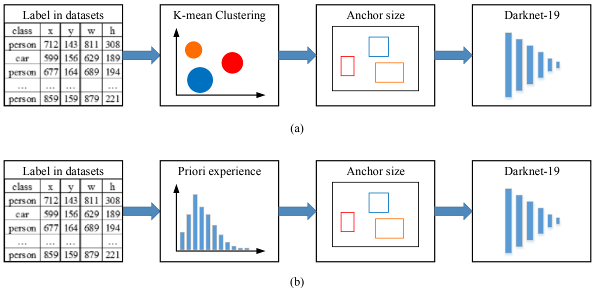

Figure 2. The comparison of different methods to extract the anchor boxes:(a) the YOLOv2 uses the K-mean clustering; (b) the proposed method

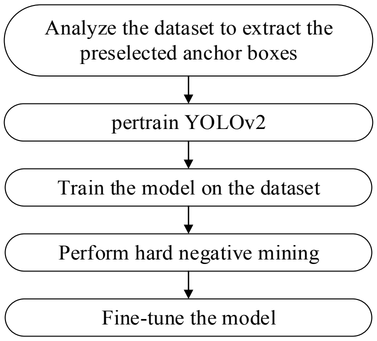
Analyze the dataset to extract the preselected anchor boxes
pertrain YOLOv2
Train the model on the dataset
Perform hard negative mining
Fine-tune the model

Figure 3. Flowchart of the training procedure of the proposed scheme

First, we analyze the label information of the pedestrian and vehicle in the KITTI dataset and the Caltech Pedestrian Dataset according to the priori knowledge. Most of the pedestrians are lean in appearance while the most vehicles are similar to the rectangle or square, which means that pedestrians' label boxes have the relationship of height> width, and the height of the vehicles' label boxes is less than or equal to their width. Based on the ratio of human, vehicle and surrounding objects and the principles of visual perspective in the driving environment, we extract the preselected anchor boxes for pedestrians, which edge ratio, height : width ≅ 3:1, and the vehicle's edge radio, height : width ≅ 2:3. Data analysis also confirms that the edge of the label boxes in the data set is also in line with this ratio. The detailed operation will be described later. We train the proposed CNN model using the Caltech Pedestrian Dataset and further use the same dataset to test the pre-trained model so as to generate hard negatives. These hard negatives are fed into the network as the second step of our training procedure. The resulting model will be further fine-tuned on the KITTI dataset. During the final fine-tuning process, we apply the multi-scale training process to further boost the performance of our model. For the whole training processes, we follow the similar end-to-end training strategy as YOLOv2. We will discuss the key steps in the proposed solution in details as follows.

B. The anchor boxes choice

According to RPN, SSD, and YOLOv2, it can be concluded that the contribution of anchoring mechanism to target detection is enormous. The YOLOv2 uses K-means clustering to select automatically the best initial boxes. However, the k-means algorithm is very sensitive to the initialization of seed points, and for the single or double classifications, the K-means clustering method is too complicated in the pedestrian detection and the vehicle detection. In this paper, we can choose the handpicked prioris. We integrate the label information of the object detection from the KITTI dataset with Caltech Pedestrian Dataset, and extract the ground truth boxes' information. Figures 4, 5 and 6 show the results of statistical analysis of pedestrian and car label ratios in different datasets, respectively, including distribution histograms and normal distributions. Figure 4 and Figure 6 show that the most probable pixel ratios between the height and width of pedestrians' boxes are in the range of [2, 3]. In Figure 4, there is a high ratio in [1.2, 1.3], because of the sitting pedestrian more in the Caltech Pedestrian Dataset and fewer marked in the KITTI Dataset. This is approximately in line with normal human characteristics. It can be seen from Figure 5 that most pixel ratios of the vehicle's height and width are between [0.36, 0.5] and [0.72, 0.75], due to different visual observations from the side of the vehicles and behind or forward the vehicles. In the actual road-driving environment, the probability distribution is in accordance with the relationship between the pedestrian and the vehicle's visual position. Figure 7 shows the label information of pedestrian and vehicle in pictures, which also verifies the previous statistical analysis results.

The YOLOv2 changes the predicted offset of YOLO to the position latitude of the predicted grid cell, limits the predicted value to 0-1, and enhances stability. This network predicts five bounding boxes for each cell in the feature map. For each bounding boxes, the model predicts five-match values (tx, ty, tw, th, to). For the detection, the proposed model predicts k anchors, and every anchor contain 5 coordinate values and two categories, so a total of k * (4+1+2) = 7*k output dimensions.

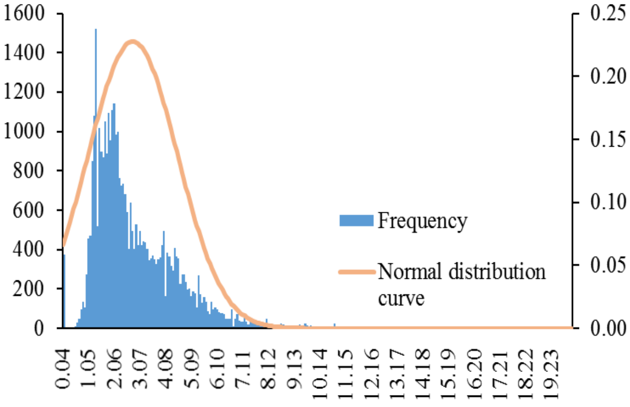

Figure 4. The histogram distribution and the normal distribution curve of the pedestrian truth boxes height / width ratio on the caltech dataseets

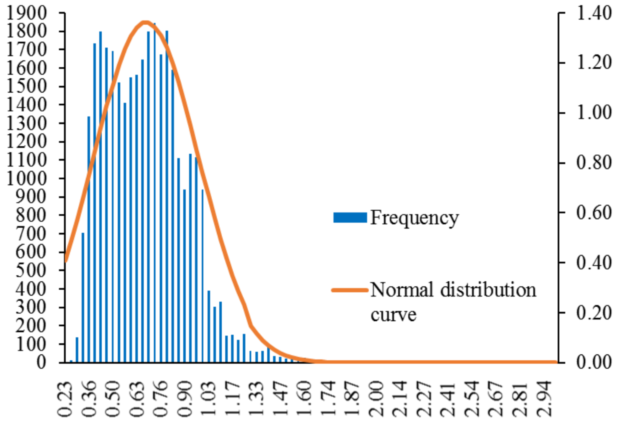

Figure 5. The histogram distribution and the normal distribution curve of the vehicle truth boxes height / width ratio on the KITTI dataseets

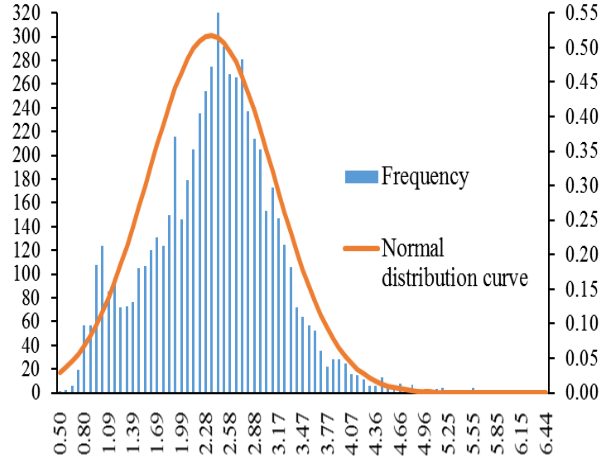

Figure 6. The histogram distribution and the normal distribution curve of the pedestrian truth boxes height / width ratio on the KITTI dataseets

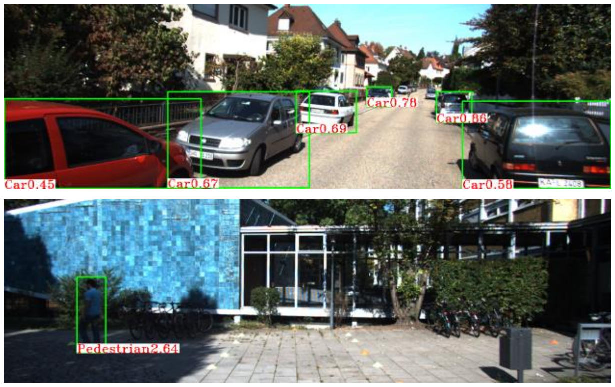

Figure 7. Label height /width ratio of pedestrian and vehicle in pictures

C. Hard Negative Mining

Hard negative mining has been proved to be an effective strategy for boosting the performance of deep learning, especially for object detection task including pedestrian and vehicle detection. This method is based on the idea which hard negatives are the regions where the network has failed to make a correct prediction. Thus, the hard negatives can be fed into the network again as a reinforcement for improved our trained model. The resulting training process will then be able to improve our model towards fewer false positives and better classification performance. In our approach, hard negatives were harvested from the pre-trained model from the first step of our training process. We then consider a region as hard negative if its IOU over the ground truth region was less than 0.5. During the hard negative training process, we explicitly add those hard negatives into the ROIs for fine-tuning the model, and balance the ratio of foreground and background to be about 1: 3, which is the same as the ratio that we use in the first step.

D. Detection tuning

According to those references, perform the following steps in order to get a more accurate model for pedestrian and vehicle region extraction and detection. So there uses a detection tuning stage to adjust the parameters of the proposed model network to extract similar pedestrian and vehicle features. In order to make the model robust to different size images, this paper uses the approach presented in the paper [14] to divide the tuning process.

During training, the model input size is changed every few rounds so that the model is robust to different sizes of images. For each number of batches, the model randomly selects a new input image size, changes the model input size, and continues to train. The training rules force the model to accommodate different input resolutions. The model is faster for small-size input processing, so the proposed models can adjust speed and accuracy as needed.

IV. EXPERIMENTAL RESULTS

In this paper, pedestrian detection and vehicle detection is evaluated using the well-known Caltech Pedestrian Dataset and the KITTI Dataset. Our pedestrian detection system is analyzed in terms of the precision and effectiveness.

A. Experiment Setup

The Dataset Introduction

The Caltech Pedestrian Dataset is the most popular and challenging dataset for pedestrian detection, which includes 350,000 annotations of 2300 unique pedestrians labeled in 250,000 frames which is about 10 hours at 30 Hz. Pedestrians are usually annotated by a full bounding box while the occluded ones have another bounding box that shows the visible area. The dataset was recorded using a monocular camera mounted in a vehicle in urban traffic. Thus, this database is well suited to vehicle-related applications or systems. The performance is evaluated with the reasonable setting which gives images which are at least 50 pixels high and whose occluded portions are less than 35% every 30 frames from set06 to set10.

The KITTI dataset is a public dataset that tests algorithms such as pedestrian detection, vehicle detection, vehicle tracking, and semantic segmentation in traffic scenarios. The object detection consists of 7481 training images and 7518 test images, comprising a total of 80256 labeled objects. All images are color and saved as the image format png (Portable Network Graphics). The dataset was recorded using two high-resolution color and grayscale video cameras equipped with a standard station wagon. The datasets are captured by driving around the mid-size city of Karlsruhe, in rural areas and on highways. Up to 15 cars and 30 pedestrians are visible per image.

Deep Learning Framework and Parameters

The proposed detection model was implemented on the popular deep learning framework Caffe. The ConvNet was trained on a single NVIDIA GeForce GTX 1080 GPUs with 8 GB of memory. The network parameters pre-trained by VOC Datasets was used to initialize the proposed model. And the model was then tuned using the KITTI dataset.

Before training, we set the number of the anchors as 5. For each anchor, we set the two values as the initial values of the anchor side lengths respectively, and he ratio between two values in each group complies with the pedestrian detection feature scale of Figure 3, 4 and 5. In the first training stage, we select Caltech pedestrian datasets to pre-train the network architecture. During training, we use stochastic gradient descent with a starting learning rate of 0.1, polynomial rate decay with a power of 4, weight decay of 0.0005 and momentum of 0.9. And we use data augmentation methods, such as random crops, rotations, and hue, saturation, and exposure shifts, like the training strategy used on the Faster R-CNN and the SSD. After pre-training, we harvest those hard negatives from the pre-trained model from the first step of our training process. We then fine tune the model using a high-resolution image on the KITTI dataset, the Caltech pedestrian datasets and those hard negatives. During training, the model input size is changed to make the model robust to images. For each image, in addition to performing the horizontal flipping, we also randomly resize it before feeding it into the network every 10 rounds. Specifically, we resize each image such that its size will be one of 480, 600, 640,1280. Before images are input to the model, the mean RGB values of the entire dataset are subtracted from each image to decrease the influence of the probability distribution on the pixel values. This training rules force the model to accommodate different input resolutions.

B. Results

In the experiments, the proposed model is evaluated on the reasonable testing set with well-known pedestrian detection models, include Faster RCNN, YOLOv2. The experimental results on the KITTI dataset are presented in the table I. Compared with YOLOv2, the proposed method is a little faster in the detection speed, and there is a certain improvement in the detection accuracy; Compared with the Faster RCNN, the proposed method is improved in the detection accuracy but faster in detection speed. The result shows the proposed detection model achieves the fastest test rate while providing an acceptable accuracy rate.

TABLE I. THE COMPARISON OF THOSE METHODS ON THE KITTI DATASET.

Item	Time cost*	FPS	Accuracy	
			Pedestrian	Vehicle
Faster RCNN	0s	0.5	65%	75 %
YOLO V2	128s	20	43.33%	59.57%
PROPOSED	0s	22	45%	61.34%

***The time of obtaining the anchor sizes**

C. *Effectiveness Analysis*

The Table I also compares those detection models in the time of obtaining the anchor sizes. The YOLOv2 used the K-mean clustering to obtain the anchor sizes, but the K-meaning clustering is easily affected by the initial k value setting, which resulting the anchor size not being able to better represent the different label feature information of categories. Especially in the case of uneven distribution of labels the distribution of the label of categories, the priori experience may save as many different label information of categories as possible. Two examples of pedestrian detection using the proposed method are shown in Figure 8. There are other factors that could degrade the performance, such as the feature extractor is confused by some labels due to that the bounding boxes are not precise.

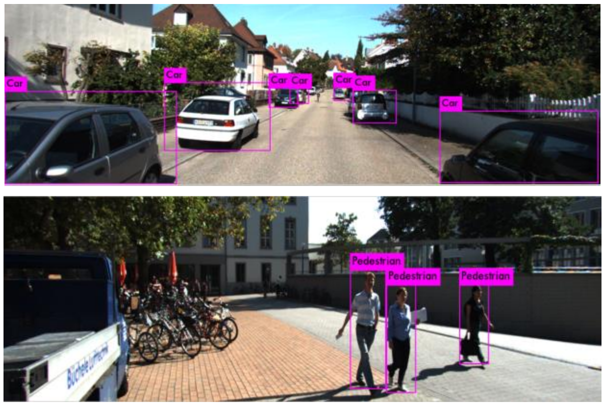

Figure 8. Two examples of the proposed detection method

V.CONCLUSION

In this paper, a new pedestrian and vehicle detection method is proposed, which is based on the YOLOv2. The feature box sizes are selected due to relatively stable height/weight ratio size of pedestrian and vehicles. This priori experiences instead of complex vision-based selection improve the detection accuracy with keeping a good detection efficiency. In the future, different ConvNets can be considered to improve the architecture; features can be integrated into the deep network to improve performance.

ACKNOWLEDGMENT

This work was supported by The National Natural Science Foundation of China (61267002, 41271362) and the Shenzhen S&T Funding with Grant No. JCYJ20160510154531467, JSGG20141020103523742, Shenzhen Engineering Lab on Autonomous Driving, and NSFC 61672512; Program for Innovative Research Team of Guilin University of Electronic Technology (IRTGUET)

REFERENCES

[1] R. Benenson, M. Omran, J. Hosang, and et al, "Ten years of pedestrian detection what have we learned?" *ECCV CVRSUAD workshop,* 2014.

[2] P. Viola and M. Jones. "Robust real-time face detection." *International Journal of CompuTer Vision*, vol.57,no.2, pp.137-154, 2004.

[3] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection." *Coference on Computer Vision and Pattern Recognition (CVPR),* 2005.

[4] P. Felzenszwalb, R.Girshick, D. McAllester, and et al, "Object detection with discriminatively trained partbased models." in *PAMI,* vol.32, pp. 1627-1645, Jul. 2010.

[5] J. Deng, W. Dong, R. Socher, and et. al, "ImageNet: A large-scale hierarchical image database." in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 248-255, 2009.

[6] M. Everingham, L. Van Gool, C. K. 1. Williams, and et. al, "The PASCAL Visual Object Classes Challenge 2011 (VOC2011)." Results [online] Available: http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html

[7] R. Girshick, J. Donahue, T. Darrell, and et. al, "Rich feature hierarchies for accurate object detection and semantic segmentation." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 580-587, 2014.

[8] K. He, X. Zhang, S. Ren, and et. al, "Spatial pyramid pooling in deep convolutional networks for visual recognition." in *Proc. Eur. Conf. Comput. Vis.*, pp. 346-361, 2014.

[9] R. Girshick, "Fast R-CNN." in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1440-1448, 2015.

[10] S. Ren, K. He, R. Girshick, and et. al, "Faster R-CNN: Towards real-time object detection with region proposal networks." in *Proc. Int. Conf. Neural Inf. Process. Syst*., pp. 91-99, 2015.

[11] J. Dai, Y. Li, K. He, and et. al, "R-FCN: Object detection via region-based fully convolutional networks." in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, pp. 379-387, 2016.

[12] J. Redmon, S. Divvala, R. Girshick, and et. al "You only look once: Unified real-time object detection" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 779-788, 2016.

[13] W. Liu, D. Anguelov, D. Erhan, and et. al, "SSD: Single shot multibox detector" in *Proc. Eur. Conf. Comput. Vis.*, pp. 21-37, 2016.

[14] R. Girshick, "Fast R-CNN." *in Proc. 2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 1440-1448, 2015, [online] Available: http://dx.doi.org/10.1109/ICCV.2015.169.

[15] C. Papageorgiou and T. Poggio. "A trainable system for object detection." *Internation Journal of Computer Vision (IJCV)*, vol.38,no.1,pp.15-33, 2000.

[16] A. Geiger, P. Lenz, C. Stiller, and et. al, "Vision meets robotics: The kitti dataset." *International Journal of Robotics Research (IJRR)*,vol.32,no.11,pp.1231-1237, 2013.

[17] P. Dollar, C. Wojek, B. Schiele, and et. al, "Pedestrian detection: An evaluation of the state of the art. " *in PAMI*, vol. 99, 2011. 3.

[18] K. He, X. Zhang, S. Ren, and et. al, "Deep residual learning for image recognition." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* pp.770-778, 2016.