2016 IEEE Intelligent Vehicles Symposium (IV)
Gothenburg, Sweden, June 19-22, 2016

A New Benchmark for Vison-Based Cyclist Detection

Xiaofei Li[1], Fabian Flohr[2,3], Yue Yang[4], Hui Xiong[5,1], Markus Braun[2,3], Shuyue Pan[4], Keqiang Li[1], Dariu M. Gavrila[2,3]

Abstract— **Significant progress has been achieved over the past decade on vision-based pedestrian detection; this has led to active pedestrian safety systems being deployed in most mid- to high-range cars on the market. Comparatively little effort has been spent on vision-based cyclist detection, especially when it concerns quantitative performance analysis on large datasets.**

We present a large-scale experimental study on cyclist detection where we examine the currently most promising object detection methods; we consider Aggregated Channel Features, Deformable Part Models and Region-based Convolutional Neural Networks. We also introduce a new method called Stereo-Proposal based Fast R-CNN (SP-FRCN) to detect cyclists based on stereo proposals and Fast R-CNN (FRCN) framework. Experiments are performed on a dataset containing 22161 annotated cyclist instances in over 30000 images, recorded from a moving vehicle in the urban traffic of Beijing. Results indicate that all the three solution families can reach top performance around 0.89 average precision on the easy case, but the performance drops gradually with the difficulty increasing. The dataset including rich annotations, stereo images and evaluation scripts (termed "Tsinghua-Daimler Cyclist Benchmark") is made public to the scientific community, to serve as a common point of reference for future research.

I. INTRODUCTION

Significant progress has been made over the past decade on improving driving safety with the development of Advanced Driver Assistance Systems (ADAS), such as pre-collision systems, crash imminent braking systems and others. In the last few years, however, extensive research interest has been focused on protecting vulnerable road users (VRUs), including pedestrians, cyclists and motorcyclists. Depending on the statistical data of WHO [1], half of the world's road traffic deaths occur among vulnerable road users. Although the likelihood of dying on the road as cyclists is less compared to pedestrians or motorcyclists as a whole, the death rates of cyclists vary by region, from 3% in Europe to 7% in Western Pacific. In some small- and mid- cities in China where cyclists appear often, even more road accidents involve cyclists. Therefore, to make cycling safer, detecting and protecting cyclists needs to be paid more attention.

Vision-based pedestrian detection has been studied for many years, but it is still a challenging problem due to the large variability in appearance, body pose, occlusion and cluttered backgrounds. Similar problems occur in the field of cyclist detection. In addition to the aforementioned problems, multiple viewpoints of cyclists bring more challenges to detect them, which is rarely taken into consideration in pedestrian detection. Cyclists can be viewed at a variety of possible orientations, which give a problem to choose the detection window size as the aspect ratio of a cyclist differs in each orientation.

As for pedestrian detection vision sensors are preferred, due to the possibility to capture a high-resolution perspective view of the scene with useful color and texture information, compared to active sensors like radar or lidar [2]. In this paper we focus on vision-based cyclist detection methods.

To deal with the issues for cyclist detection as mentioned above, there is a need for a large-scale cyclist dataset to cover the great variability of cyclists. Therefore, as the main contribution of this paper, a practical and richly annotated cyclist dataset for training and evaluating cyclist detectors is introduced. We made this dataset publicly available for non-commercial purposes to encourage research and benchmarking[1]. It contains a large number of cyclists varying widely in appearance, pose, scale, occlusion and viewpoint, and provides a common point of reference for cyclist detection and evaluation. The dataset also includes vehicle state, camera information and corresponding stereo image pairs which were recorded from a vehicle-mounted stereo vision camera.

The second contribution is a benchmark of several promising object detection methods evaluated on the new cyclist detection dataset, including Aggregated Channel Features (ACF) [3], Deformable Part Models (DPM) [4] and Region-based Convolutional Neural Networks (R-CNN) [5]. All the selected methods are obtained directly from the original authors or re-implemented with adjustment for cyclist detection. In addition to the commonly used Fast R-CNN (FRCN) method [6], we also introduced a new method called SP-FRCN, which utilizes the power of FRCN combined with stereo proposals extracted from the stixel world.

II. PREVIOUS WORK

Challenging datasets have promoted technological progress in computer vision. As a hot topic, there are already some publicly available pedestrian datasets, such as the INRIA [7], Caltech [8] and Daimler [9] [10] pedestrian detection datasets, which promote the development of pedestrian detection. On these datasets, cyclists are excluded or ignored because of the similar appearance between

[1]State Key Laboratory of Automotive Safety and Energy, Tsinghua University, China
[2]Environment Perception Department, Daimler R&D, Ulm, Germany
[3]Intelligent Systems Laboratory, Univ. of Amsterdam, The Netherlands
[4]Driver Assistance and Chassis Systems, Daimler Greater China Ltd., Beijing, China
[5]School of Software, Beihang University, China

[1]This dataset is available for non-commercial research purposes. Follow the links from http://www.gavrila.net or contact the second author.

pedestrians and persons riding a bicycle. Although cyclists are often encountered in traffic accidents, especially in some developing countries, there is no challenging cyclist dataset publicly available yet, except the KITTI object detection benchmark [11]. However, there are very limited cyclist instances (no more than 2000) in the training set, which might not be sufficient for cyclist detection and evaluation.

Vision-based pedestrian detection has been extensively investigated over the last decade, the readers are referred to some general surveys, such as [2], [10] and [12]. As opposed to pedestrian detection, very limited work has been undertaken in the domain of vision-based cyclist detection, although similar techniques are used for cyclist detection. Li [13] used HOG-LP features and linear SVM classifier to detect crossing cyclists, with the purpose of optimizing the time-consuming steps of HOG feature extraction. Chen [14] proposed a part-based bicycle and motorcycle detection for nighttime environments integrating appearance-based features and edge-based features. Cho [15] defined a mixture model of multiple viewpoints to detect cyclists, which was based on part-based representation, HOG feature and SVM. In [16], an integral feature based detector was applied to filter out most of the negative windows, then the remaining potential windows were classified into cyclist or non-cyclist windows by three pre-learned view-specific detectors. In order to handle the multi-view problem of cyclists, the work proposed in [17] divided the cyclists into subcategories based on cyclists' orientation. For each orientation bin, they built a cascaded detector with HOG features.

We note that little effort has been spent on quantitative performance analysis for cyclist detection in the aforementioned literatures. So in the reminder of this paper, we benchmark several representative object detection methods in the new dataset to show their detection performance quantitatively.

III. SELECTED CYCLIST DETECTION METHODS

In recent years, many methods have been designed to address object detection tasks. In the field of pedestrian detection, three families are representative: decision forests, DPM variants and deep networks [12]. In this paper, the detectors based on ACF [3], DPM [4] and R-CNN [5] are revisited as the representative methods for each family to detect cyclists and evaluate the new cyclist benchmark.

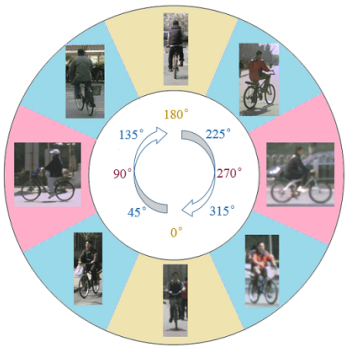


Fig. 1. Cyclists are divided into three classes: wide (red), intermediate (blue) and narrow (brown).

A. Division of Positive Samples

In order to deal with the high intra-class variation of cyclists in different views, we divide the cyclist samples into three classes: narrow, intermediate and wide. Unlike the work in [17] dividing positive samples into eight equidistant orientation bins based on the ground truth of orientations, we divide them based on the aspect ratio of bikes, which is more discriminative than the aspect ratio of cyclists for different views.

Without exhaustive search, we divide the positive samples into three classes based on two empirical break points: 0.75 and 1.75, which means a cyclist belongs to

$$\begin{cases} wide, & if\ h/w \leqq 0.75 \\ intermediate, & if\ 0.75 < h/w \leqq 1.75. \\ narrow, & if\ 1.75 < h/w \end{cases} \quad (1)$$

Here h and w are the height and width of a bike. Different classes of cyclists are shown in Fig. 1. Note that the break points are chose empirically without exhaustive search. Of course, appreciate parameters might improve the final performance slightly, which won't be discussed in this work.

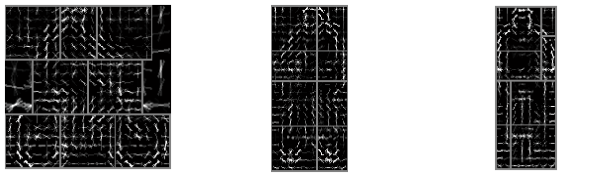
B. Aggregated Channel Features

The channel features based detectors proposed by Dollar et al. [3] are utilized to detect cyclists in this work because the methods are conceptually straightforward and efficient. Specifically for our work in this paper, original aggregated channel features (ACF) and locally decorrelated channel features (LDCF) variant [18] are employed. Take the latter for example, given an input image, LDCF computes several feature channels and removes correlations in local neighborhoods of feature channels, where each channel is a per-pixel feature map such that output pixels are computed from corresponding patches of input pixels.

We choose model window sizes without padding [50,18], [50,24] and [50,50] for three classes respectively. Each aspect ratio of the chosen windows is close to the average aspect ratio of each class. And the window sizes with padding are set as [64,32], [64,48] and [64,64] respectively for training cascaded detectors. During the test phase, each of the three detectors is applied to detect cyclists separately, resulting in multiple overlapping detections for each object. Therefore, we assemble three detection outputs directly and use non-maximum suppression to get the final detection results.

C. Deformable part model

The deformable part model (DPM) proposed by Felzenszwalb et al. [4] is one of the most popular object detection methods, especially in dealing with appearance variations and partly occlusions, due to the fine-grained spatial relationships between one root and many parts. DPM detects objects by figuring out all possible locations for distinguished root filters and finding the best configuration of the remaining parts for each root filter.



(a) wide view (b) intermediate view (c) narrow view

Fig. 2. HOG feature representations of trained cyclist models for three classes learned from the new cyclist dataset.

Before training, we divide the positive samples into three views manually, different from the original DPM which clusters them into three groups based on the aspect ratio automatically. During training, star-structured models and 32-dimensional HOG features are chosen to train the detectors. Besides, the number of bins, the mixture model and other configurations are kept just as the original DPM method. To demonstrate the effectiveness of the positive sample division method, HOG feature representations of a trained cyclist detector from three viewpoints are visualized in Fig. 2.

D. Region-based Convolutional Neural Network

Recent advances in object detection are driven by the success of R-CNN [5], which involves a category-independent region proposal method to extract the set of candidate detections, a large convolutional neural network to extract object feature vectors and a linear SVM to classify object classes. As an upgraded version of R-CNN, FRCN [6] is deployed to detect cyclists in this paper. Unlike the original R-CNN, FRCN trains a single-stage multi-task loss network. The inputs of the network are the whole image and a set of proposals. After several convolutional and max pooling layers, a region of interest pooling layer extracts a fixed-length feature vector from the feature map. Finally two sibling output layers, which produce classification probability and bounding box regression, are connected after a sequence of fully connected layers.

Unlike the original method, which either uses Selective Search (FRCN [6]) or a Region Proposal Network (Faster R-CNN [19]) for extracting relevant proposals, our new method SP-FRCN utilizes stereo data for proposal generation based on the *stixel* representation [20]. Let a stixel be described in the car coordinate system with its lateral, longitudinal position and height $[x_c, z_c, h_c]$. For each stixel in the range $z_c \in [4, 100]$m and $h_c \in [1.2, 2.4]$m, we generate three proposals with different aspects (1:2, 2:3, 1:1), the same center position as the stixel and a proposal height $h_p = h_c$. For stixels with $h_c > 2.4$m, we sample the proposal height $h_p \in [1.2, 2.4]$m with a step size of 0.3m. By this we get 18/4002/815 (min/max/average) proposals per image. Fig. 3 shows an image of filtered stixels that are used for our proposal generation.

IV. DATASET OVERVIEW

Approximately 6 recording hours were collected from a vehicle-mounted stereo vision camera (image resolution of 2048×1024 pixels, baseline of 20 cm) at 25 Hz driving through regular urban traffic in 5 different days. The videos were collected in the northern city of Beijing chosen for their relatively high concentration of cyclists and pedestrians: Haidian District and Chaoyang District. Besides the images, IMU information, including velocity, longitudinal acceleration and yaw rate, was captured concurrently to offer useful vehicle information for different application tasks.

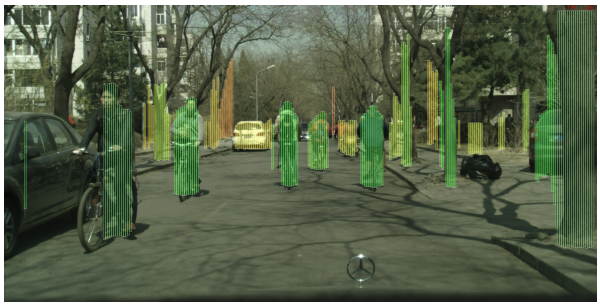


Fig. 3. Remaining stixels after applying the defined constraints. For each stixel, multiple proposals with different aspects are generated to be used within SP-FRCN.

We annotated about 14674 frames from more than 5 million images for a total of 32361 labeled vulnerable road users (VRUs), including cyclists, pedestrians, tri-cyclists and motor-cyclists etc. Fig. 4 shows an excerpt from the new dataset. For all the manually labeled images, we divided them into two clusters: partly labeled and fully labeled.

In partly labeled images, only ideal cyclists were annotated with two tight bounding boxes, indicating the full extent of the rider and the bike respectively. Cyclists lower than 60 pixels, occluded or truncated more than 10% or motion blurred were ignored. Pedestrians, tri-cyclists and motor-cyclists were also ignored. The images were annotated every 10 frames in the partly labeled set.

In fully labeled images, all VRUs were annotated as long as they were higher than 20 pixels, not occluded more than 80% and not truncated more than 50%. For cyclists and other riders, two tight bounding boxes were annotated, indicating the full extent of the rider and the bike respectively. For pedestrians, only one bounding box indicating the extent of pedestrians was annotated. Three discrete occluded levels were tagged for all objects, which mean no occlusion (occlusion $\leq$ 10%), partially occluded (10% $<$ occlusion $\leq$ 40%) and heavy occluded (40% $<$ occlusion $\leq$ 80%). Besides, truncated level was also tagged for the objects which were not truncated by image borders more than 50%. In the fully labeled set annotation was done every 5 frames.

In order to support different research groups to compare their methods, we split the dataset into training and test set. All of the partly labeled dataset was utilized as training data, termed training set 1, and part of the fully labeled dataset were extracted as training set 2. Additional 1000 images without any VRUs were extracted to supplement the training dataset, termed non-VRU set. Left fully labeled dataset was used as test set. Table 1 shows the detailed statistics.

In order to understand the new cyclist dataset well, we give a detailed analysis of the distribution of cyclist scales

Fig. 4. Overview of the new cyclist detection dataset. (a) Cyclist samples; (b) Test images with annotations: green, blue and yellow bounding boxes indicate cyclists, pedestrians and other riders respectively.

TABLE I
STATISTICS OF TSINGHUA-DAIMLER CYCLIST BENCHMARK

	Training			Test set	Total
	Set 1	Set 2	Non-VRU		
Total Frames	9741	5095	1000	14570	30406
Labeled Frames	9741	1019	1000	2914	14674
Total BBs	16202	3016	0	13143	32361
Cyclist BBs	16202	1301	0	4658	22161
Pedestrian BBs	0	1539	0	7380	8919
Other rider BBs	0	176	0	1105	1281

for the partly and fully labeled dataset respectively. In Fig. 5, we histogram the heights and height/width aspect ratios of cyclist bounding boxes from partly and fully labeled dataset. Note that we histogram the height of cyclists distribution using logarithmic sized bins, which show most of the labeled cyclist are concentrating between 30 and 500 pixels. Compared to the aspect ratio of pedestrians in [8], the distribution of cyclists' aspect ratio (a cyclist contains a bike and a rider) is more decentralized, due to multiple views of cyclists. Besides, on the fully labeled dataset, 87.91% cyclists are almost fully visible, 9.21% cyclists are partially occluded, and the left 2.88% are heavily occluded. Only 1.23% cyclists are partly truncated by image borders.

V. EXPERIMENT

In this section, evaluation parameters and detailed settings of different methods are described to allow reproducibility of the results. The performance evaluation of different detectors on the new cyclist test dataset is compared and discussed.

A. Evaluation protocol

For evaluation, the well-established methodology used in the PASCAL object detection challenges [21] is utilized to show the relationship between precision and recall rate. Meanwhile, the average precision (AP) is used to summarize the performance of precision/recall curve. To assign the output detections to ground truth objects, the PASCAL measure is employed, which states that the area of IoU overlap must exceed the threshold of 0.5. To evaluate performance on various subsets of the new cyclist dataset, similar to [11], we define three difficulty levels as follows:

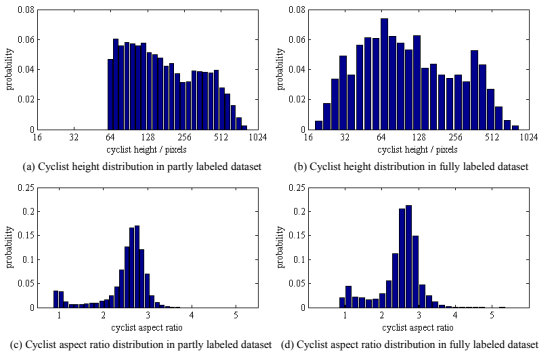


Fig. 5. Cyclist scale distribution of the new cyclist dataset.

- Easy: cyclists with bounding boxes higher than 60 pixels and fully visible.
- Moderate: cyclists with bounding boxes higher than 45 pixels and less than 40% occlusion.
- Hard: cyclists with bounding boxes higher than 30 pixels and less than 80% occlusion.

During evaluation on a subset, the objects not included in the subset are ignored instead of discarded directly, which need not to be matched with detections. Besides, we also want to evaluate the ability of the detectors to distinguish cyclists from pedestrians or other riders. Therefore we also compare the detection performance between ignoring and discarding pedestrians and other riders in the following experimental section.

B. Parameter configuration

During training ACF and DPM detectors, we extracted cyclists from training set 1 and flipped them to augment the positive samples, and extracted negative samples from both training set 2 and non-VRU set. We divided the positive samples into three classes. As a result, 1254, 4863 and 10085 positive samples were classified into the wide, intermediate and narrow classes respectively. For mining hard negative samples from training set 2 and non-VRU set, we set IoU overlap threshold as 0.3. The latest version of Piotr's Computer Vision MATLAB Toolbox [3] was applied in this work to train three view-based detectors separately with the

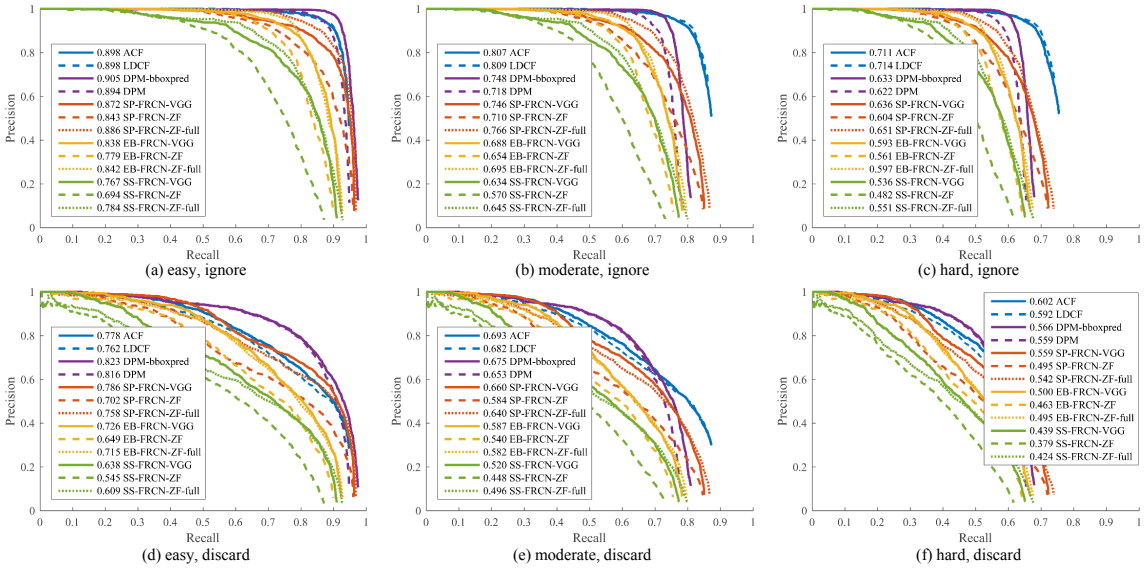


Fig. 6. Precision versus recall curves of various detectors shown for various subsets of the new cyclist test dataset. The AP is listed before the name of each method. When evaluating cyclist detectors, (a)∼(c) ignore pedestrians and other riders, (d)∼(f) discard pedestrians and other riders.

same parameters in the original application [18]. For the DPM detector, the voc-release5 code [4] was applied with almost the same parameters as the original work, except the viewpoint clustering and bootstrapping images, which were described in the previous section.

For training our SP-FRCN method, the open source of Fast RCNN [6] with pre-trained ZF-nets and VGG-nets was applied in this work. Selective Search (SS) [22] and Edge Boxes (EB) [23] methods were also considered for comparing different proposal methods. For fine-tuning the deep networks, its final sibling layers are adapted to this task. Each SGD mini-batch was constructed from 2 images. The length of the shortest input image side was limited to 650 pixels and the longest image side was capped at 1300 pixels for VGG-nets, due to limited graphics memory. For ZF-nets, we also considered another big input scale: 1024 pixels for the shortest image side and 2048 pixels for the longest image side. The first image of a batch was chosen from training set 1, and the second images was chosen from training set 2 or non-VRU set. Both of the images were chosen uniformly at random from corresponding dataset. We used mini-batches of size 128, sampling positive samples (max 25% of batch size) from the images of training set 1 and 2 with a minimum overlap of 0.5 with a ground-truth bounding box. The left negative samples were sampled from all the images of the training dataset with a maximum overlap of 0.5: negatives around positive samples were extracted from training set 1 and 2; hard negatives and additional normal negatives were extracted from the images of training set 2 and non-VRU set. We did bootstrapping every 10000 iterations to mine hard negative samples. We used a learning rate of 0.001 for 30000 iterations, and 0.0001 for the next 10000 iterations. Other network's configurations and parameters were the same as in the original paper [6]. During the test phase, all the detectors mentioned above deployed a greedy fashion of non-maximum suppression to suppress bounding boxes with lower scores, just like the methods used in ACF [3].

C. Detection performance

Fig. 6 illustrates the overall detection performance of the selected detectors. In the easy subset, all the three solution families can achieve a high AP when ignoring pedestrians and other riders, showed in Fig. 6 (a). Among them, DPM with bounding boxes regression slightly outperforms the other detectors, achieving a 0.905 AP. Fig. 6 (b) and (c) show that, with the subset becoming harder, APs of all detectors decrease gradually, where cyclists are at lower resolution and under partial occlusion. It can be observed that ACF and LDCF outperform other methods in the moderate and hard sets. If we discard pedestrians and other riders directly, APs of all detectors drop significantly, seen in Fig. 6 (d)∼(f), which illustrates all detectors can't distinguish cyclists from pedestrians and other riders perfectly if we don't train pedestrians and other riders against cyclists specifically.

To our surprise, FRCN-based detectors fall behind compared to ACF or DPM methods under almost all conditions. This may be attributed to two reasons: the input image size of VGG-nets is limited to small one (650×1300); the capacity of ZF-nets is not as strong as VGG-nets. We also train another ZF-nets with full-size input images, which can achieve better results than VGG-nets and ZF-nets with downscaled input images. Therefore, we guess a FRCN-based detector with bigger input images and deeper networks may achieve better results. When comparing FRCN-based detectors, SP-FRCN outperforms SS-FRCN and EB-FRCN significantly, which illustrates the benefit of our stereo proposal method for cyclist detection.

D. Evaluation of cyclist proposal methods

In order to compare the proposal methods used in the FRCN framework directly, we compute the recall of different

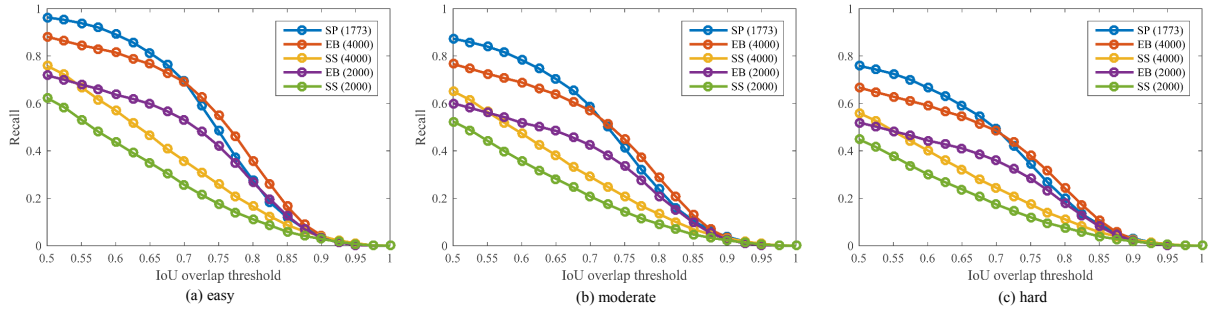


Fig. 7. Recall versus IoU threshold of different proposal methods shown for various subsets of the new cyclist test dataset. SP, EB and SS indicate stereo proposal, edge boxes and selective search methods respectively. The number of proposals is also listed following each method's name.

proposals at different IoU ratios with ground truth bounding boxes, shown in Fig. 7. SS and EB methods are utilized with default parameters. The N proposals are the top-N ranked ones based on their confidences.We consider different numbers (2000 and 4000) for SS and EB.

The plots show that the stereo proposal method behaves gracefully on the new cyclist test dataset. It can be seen that with only 1773 proposals, we achieve 96.29%, 87.34% and 75.85% recall in the easy, moderate and hard subsets respectively when IoU overlap is 0.5. With approximate proposal number, we can see the stereo proposal method outperforms other methods significantly. This explains why the stereo proposal method has a good ultimate detection performance when using fewest proposals.

VI. CONCLUSION AND FUTURE WORK

In this paper, a large and richly annotated cyclist stereo benchmark for training and evaluating cyclist detectors is introduced. State-of-the-art object detection methods are carefully benchmarked on the task of cyclist detection using the new dataset. We also introduce a new method called SP-FRCN, which utilizes the power of FRCN combined with stereo proposals extracted from the stixel world.

For cyclists from the easy subset, all the three solution families including ACF, DPM and FRCN can reach top performance around 0.89 AP in the easy subset, while SP-FRCN outperforms the state-of-the-art FRCN-based significantly. For cyclist proposal methods, stereo proposal method outperforms Selective Search and Edge Boxes, even with less proposed bounding boxes, which shows the importance of stereo information for cyclist detection.

However, there is still a big room to improve on the moderate and hard subset, where cyclists are at lower resolution and under partial occlusion. Further research addressing detection at smaller and partially occluded cyclists is crucial, which could be enhanced by multiple cues [24] in the detection task. In order to make cycling safer, cyclists' orientation estimation [25] and path prediction [26] could help to improve risk assessment. Therefore, we are planning to extend our cyclist benchmark to explore cyclists' orientation and path prediction in the future work.

REFERENCES

[1] World Health Organization. WHO global status report on road safety. *World Health Organization*, 2015.

[2] D. Geronimo, A. M. Lopez, A. D. Sappa, T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. on PAMI*, 32(7): 1239-1258, 2010.

[3] P. Dollar, R. Appel, S. Belongie, P. Perona. Fast feature pyramids for object detection. *IEEE Trans. on PAMI*, 36 (8): 1532-1545, 2014.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. on PAMI*, 32(9): 1627-1645, 2010.

[5] R. Girshick, J. Donahue, T. Darrell, J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, 2014.

[6] R. B. Girshick. Fast R-CNN. *arXiv preprint arXiv:1504.08083*, 2015.

[7] N. Dalal, B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005.

[8] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. on PAMI*, 34 (4): 743761, 2012.

[9] C. G. Keller, M. Enzweiler, D. M. Gavrila. A new benchmark for stereo-based pedestrian detection. *IV*, 2011.

[10] M. Enzweiler, D. M. Gavrila. Monocular pedestrian detection: survey and experiments. *IEEE Trans. on PAMI*, 31(12): 2179-2195, 2009.

[11] A. Geiger, P. Lenz, R. Urtasun. Are we ready for autonomous driving? The KITTI Vision Benchmark Suite. *CVPR*, 2012.

[12] R. Benenson, M. Omran, J. Hosang, B. Schiele. Ten years of pedestrian detection, what have we learned? *ECCV*, 2014.

[13] T. Li, X. Cao, Y. Xu. An effective crossing cyclist detection on a moving vehicle. *Proc. of Intelligent Control and Automation*, 2010.

[14] H. Chen, C. Lin, W. Wu, Y. Chan, L. Fu, P. Hsiao. Integrating appearance and edge features for on-road bicycle and motorcycle detection in the nighttime. *ITSC*, 2014.

[15] H. Cho, P. E. Rybski, W. Zhang. Vision-based bicyclist detection and tracking for intelligent vehicles. *IV*, 2010.

[16] K. Yang, C. Liu, J. Zheng, L. Christopher, Y. Chen. Bicyclist detection in large scale naturalistic driving video. *ITSC*, 2014.

[17] W. Tian, M. Lauer. Fast Cyclist Detection by Cascaded Detector and Geometric Constraint. *ITSC*, 2015.

[18] W. Nam, P. Dollar, J. H. Han. Local decorrelation for improved pedestrian detection. *NIPS*, 2014.

[19] S. Ren, K. He, R. Girshick, J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *NIPS*, 2015.

[20] M. Enzweiler, M. Hummel, D. Pfeiffer, U. Franke. Efficient stixel-based object recognition. *IV*, 2012.

[21] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman. The Pascal Visual Object Classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2): 303-338, 2010.

[22] J. Uijlings, K. van de Sande, T. Gevers, A. Smeulders. Selective search for object recognition. *Int. J. Comput. Vis.*, 104(2): 154-171, 2013.

[23] C. L. Zitnick, P. Dollar. Edge boxes: Locating object proposals from edges. *ECCV*, 2014.

[24] M. Enzweiler and D. M. Gavrila. A Multi-Level Mixture-of-Experts Framework for Pedestrian Classification. *IEEE Trans. on Image Processing*, 20(10): 2967-2979, 2011.

[25] F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, D. M. Gavrila. A probabilistic framework for joint pedestrian head and body orientation estimation. *IEEE Trans. Intell. Transp. Syst.*, 16(4): 1872-1882, 2015.

[26] J. F. P. Kooij, N. Schneider, F. Flohr, D. M. Gavrila. Context-based pedestrian path prediction. *ECCV*, 2014.