Real-time Large Scale Traffic Sign Detection

Aleksej Avramović, *Member, IEEE,* Domen Tabernik, and Danijel Skočaj, *Member, IEEE*

***Abstract* — Automatic traffic sign detection and recognition has achieved good results using convolutional neural networks. Novel architectures are still being proposed in order to improve accuracy of detection and segmentation of traffic sings. In this paper, we are examining the possibility for traffic sign detection and recognition in real-time. For that purpose, we employed a novel YOLOv3 architecture, which has been proven to be fast and accurate method for object detection. It was shown that real-time detection can be achieved, even on HD images, with mAP above 88%.**

***Keywords* — Deep learning, traffic sign detection, real-time, traffic sign recognition, YOLO.**

I. Introduction

MODERN vehicles are usually equipped with number of sensors, and both front and rear video cameras. Their original purpose was parking and driving assistance, both of which was done by visual or sound alarms. Front vehicle cameras can be used for warning of nearby pedestrians, bicycles and other vehicles as well as for detection of other kind of obstacles. With current advances in technology and car industry it is reasonable to expect that car computers will be more powerful and able to provide more assistance, also regarding traffic sign detection and recognition and assessment of road conditions. In order to be able to provide the driver with useful data fast enough, car computers should be able to process the data in real time. Thus, it is important that traffic signs can be recognized in real-time, so driver is informed instantly.

An important issue of both automatic traffic sign detection (TSD) and traffic sign recognition (TSR) is management of traffic sign inventory, which is important for maintenance of traffic signs on road and highways as well for safety of participants in traffic. Automatic TSD/TSR is crucial for both driver assistance/autonomous driving and management of traffic sign inventory for road maintenance. Considering the previous facts, automatic traffic sign detection and recognition gained much deserved attention in image processing and computer vision community. A significant number of tailored algorithms designed for traffic sign detection and recognition was proposed [5]-[8].

Recent breakthrough in the field of machine learning, which relies on deep learning techniques, allowed the implementation efficient object recognition algorithms. Also, this approach was successfully implemented for traffic sign detection and recognition. Convolutional neural networks (CNN's) are the state-of-the-art methods, which give the best result in object detection and recognition, therefore notable performance is achieved in the task of traffic sign detection (TSD) and traffic sign recognition (TSR) [1]-[4]. A various deep learning techniques were successfully implemented for sign recognition. In [9]-[11] CNN's were used to extract feature vectors to perform traffic sign recognition. Also, several other network learning methods were implemented to further improve accuracy.

There are traffic sign databases available online, that were gathered and annotated by different research groups in order to create a benchmark suitable for both detection and recognition tasks. some of the mostly used are: German Traffic Sign Recognition Benchmark [2] including 43 classes for recognition only; Belgium Traffic Sing [12] dataset with 62 classes for both detection and recognition; the extended MASTIF database [13][14] with 31 classes; the Tsinghua-Tencent 100K dataset [4] with 45 classes. However, these databases usually have limited number of different traffic sing. Thus, it is difficult to fairly evaluate and compare different detection and recognition approaches. There are also several private traffic sing databases, some of them with more than hundred classes.

In this research, we evaluated the performance of real time traffic sing detection on large scale dataset with 200 different classes, using deep learning approach. For this purpose, the latest YOLO detector was used, since it was proven to have good performance in general object detection as well as short detection time. Different backbone network architectures and values of hyper parameters were used to evaluate detection/recognition speed and accuracy. During detection and recognition of traffic signs, they can be clearly visible from the near distances in which case they are feature rich as easy to detect. On large distances and in cases when they have elongated shapes, discriminative features may be missing. We are particularly interested how accurate recognition is, regarding the shape of traffic sign instances.

This research is partially supported by the program of academic specialization, granted by Ministry of Science and Technology of the Republic of Srpska and by the Ministry of Science and Technology of the Republic of Srpska under contract 19/6-020/961-37/15.

Aleksej Avramović is with the Faculty of Electrical Engineering, University of Banja Luka, Patre 5, 78000 Banja Luka, Bosnia and Herzegovina (e mail: aleksej.avramovic@etf.unibl.org).

Domen Tabernik is with the Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana, Slovenia (e mail: domen.tabernik@fri.uni-lj.si).

Danijel Skočaj is with the Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana, Slovenia (e mail: danijel.skocaj@fri.uni-lj.si).

This paper is further organized as follows. In Section II details on network architecture for real time traffic sign detection are given as well on the database which is used for the experiments. Section III gives details on methodology and discussion on experimental results and Section IV gives conclusion remarks.

II. Real-time traffic sign detection

A. Object Detection Network

"You only look once" (YOLO) is object detection method with deep neural network as a backbone, which is designed for fast and accurate general object detection. It is designed to simultaneously predicts multiple bounding boxes and class probabilities for those boxes. Next, YOLO takes the whole image into consideration during training, so it is able to consider contextual information about objects. Nevertheless, since it is able to quickly identify objects, it is unable to precisely localize them, especially if objects are small. Thus, we want to examine the trade-off between traffic sign detection speed and traffic sign recognition accuracy.

YOLO takes input image and divide it into rectangular grids, after which each grid predicts a number of bounding boxes. For each bounding box a confidence is calculated to estimate probability that box contains an object. Next, for each grid that contains an object conditional class probability is calculated to estimate the class to which that object belongs. During the testing, conditional class probabilities and box confidences are combined to encode both probability that class is appearing in the box and how well the box fits the object [15]. The first version of YOLO included 24 convolutional layers for feature extraction and 2 fully connected layers for probabilities prediction. Improvements described in [16] increased detection accuracy on benchmark object-detection databases. Among all, authors used 30-layer architecture, introduced a multiscale training method to jointly train object detection and classification and introduced identity mapping by concatenating the feature maps from different layers. Interesting properties of YOLOv2 was ability to train on large scale databases and more accurate detection of small objects.

Finally, YOLOv3 presented in [17] made additional adjustments. First of all, an underlying network consist of 106 layers, which makes detection little bit slower comparing to previous version. Architecture includes residual blocks, skip connections and upsampling. Next, YOLOv3 makes detections at three different scales, from three different places in the network and, a larger number of boundary boxes is considered. The simpler version, called Tiny YOLO with total number of 22 layers, may also be used for faster detection, but with the cost of reduced detection accuracy.

B. Large Scale Traffic Sign Database

The used data is based on [18] and extended with a few additional images. It was collected by private company for the purpose of maintaining inventory of traffic signs. A large number of images were captured from the RGB camera mounted on the vehicle in both urban and rural areas. Images that contain at least one traffic sign instance were kept in database and carefully annotated with bounding boxes. A majority of images' size is 1920x1080 pixels, while a minority has less resolution. Annotations are done only for planar traffic sings. Furthermore, signs smaller than 30 pixels are discarded as well as classes with less than 20 samples. This way, a sufficient number of features per class is ensured for network evaluation. In total, over 5000 pictures with more than 13k annotations, distributed in around 200 class was included in database. Training/test split is done to ensure that at least 25% of traffic sing instances from one class must appear in the test set.

Additional synthetic traffic sign instances were created to expand existing database [18]. Data augmentation included a random mapping of tightly cropped traffic signs into real-world images which were without traffic signs previously. Before mapping, a slight distortion in appearance of traffic signs were done. First type of distortions included changes in shape and scale, while second type included variations in brightness and contrast. These distortions were done according the distributions of training set's geometry and variability in order to make new traffic sign instances as realistic as possible. This way, an original database was expanded with more than 8k synthetic images containing more than 30k traffic sign instances. Each class contains at least 200 instances, instead of 20 in original database.

Different traffic sign classes belong to different categories, that have various shapes and appearances in order to provide different kind of warnings and information. Category I are ***warning*** signs that are triangular shape with red frame. Category II are ***prohibitory*** and ***mandatory*** signs that are usually round, with white background and red frame, or with blue background and without frame. Also, this category includes stop sign. Category III are ***informative*** signs which can be round or rectangular with various colors and graphics. ***Supplementary*** signs belong to category IV and they are usually rectangular shape below the corresponding sign. Category VI contain bumpers, vertical obstacles and poles. They are usually rectangular with different color stripes. Category VII includes road signs with different arrow directions and titles, while category X includes the traffic mirrors.

III. Experiments

A. Methodology

Detection with YOLOv3 is done using publicly available implementation [17] based on Darknet network. Each model is initialized with weights of network pre-trained on ImageNet database, also available at [17]. Number of filters in the last layer of YOLOv3/Tiny YOLO were modified so 200 classes can be detected. Also, anchors were calculated from training set and used during the training process. Input image size is set to 608x608 pixels, with randomized image resizing during the training. In each case batch size was 64 with 2 images in subbatch for YOLOv3 and 8 images in subbatch for Tiny

YOLO. For YOLOv3, learning rate is set to 0.001 and decreased after each 15000 iterations, while in Tiny YOLO learning rate decreasing is done after 10000 iterations. Approximately, 400 epochs were used to train both models. Each network is trained on computer with two 1080ti GPU's. Experiments are done on both basic and database extended with data augmentation, as described in Section II.

B. Results

To evaluate the possibility of real-time TSD/TSR we trained and tested different YOLO networks and reported both accuracy and time consumption. The accuracy is given by standard mean averaged precision (mAP^{50}), as it was defined in [19]. We also reported precision, recall, F1-score and averaged intersection of union (IoU). IoU estimates the overlap between detection and ground truth boundary box and if it is above the predefined threshold, detection is considered as true. Usually, threshold value of 50 is used as in [19], which means that object is detected if overlaps more than 50% with its ground truth boundary box. Computational and time consumption is given by averaged “billions of FLOPS” (BFLOPS) and time required to process one single image. Results are summarized in Table I.

TABLE I: RESULTS AND DETECTION TIME OF TSD/TSR USING YOLO DETECTION METHOD AND 1080TI GPU

average	without augmentation		with augmentation	
	YOLOv3	Tiny YOLO	YOLOv3	Tiny YOLO
Precision	0.87	0.83	0.87	0.82
Recall	0.80	0.61	0.84	0.58
F1-score	0.83	0.70	0.85	0.68
BFLOPS	143	13	143	13
time	0.036	0.012	0.036	0.012
IoU	75.1	70.6	76.3	70.0
mAP^{50}	84.1	72.0	88.1	71.3

As expected the highest accuracy is achieved with YOLOv3 trained with expended database. However, since recall is 0.84, it's worthy to try further improvements. Also, it's interesting to analyze averaged precision per each class, which is given in Fig. 1. We can notice that majority of classes have averaged precision (AP^{50}) above 80%, in this case exactly 161 of them. For 62 classes AP^{50} is equal to 100%, but for some of them detection is poor with AP^{50} less than 40%.

A several interesting cases of traffic sign detection using YOLOv3 are given in Fig. 2. The left column gives detection results for several testing images, while the right column is the corresponding ground truth annotation. Given images are cropped from the original in order to expose the part with the traffic signs.

TABLE II: PER CATEGORY MAP^{50} FOR YOLOV3 AND TINY YOLO, ON EXTENDED DATASET.

	I	II	III	IV	VI	VII	X
YOLOv3	95.1	90.6	89.2	88.9	74.5	49.0	81.2
Tiny YOLO	73.2	76.1	77.2	64.2	58.6	24.2	65.0

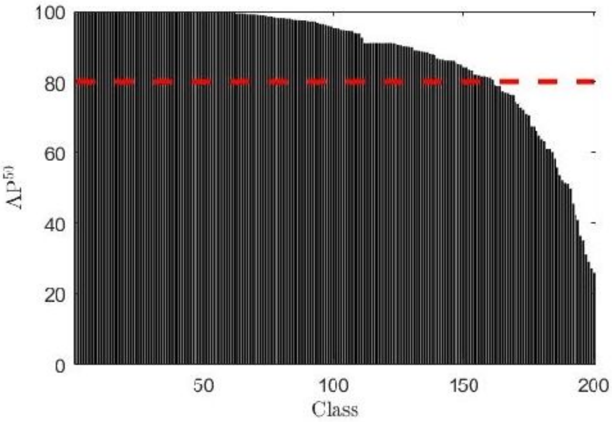

Fig. 1. Per class averaged precision AP^{50} for YOLOv3 on extended database.

C. Discussion

From the results given in Table I we can notice that the average detection time per image is little less than 40 ms for YOLOv3, which means that approximately 25 images can be processed in one second. Considering that images are mostly in HD resolution, even the borderline of real-time frame rate is sufficient. Furthermore, Tiny YOLO has three times shorter detection time. Of course, high detection speed is achieved with powerful computational capabilities of GPU's with multiple CUDA cores.

As expected YOLOv3 is much more accurate comparing to Tiny YOLO, also the network trained with the expanded dataset improved mAP^{50} for ~4%. Although detection is very fast, it is worth to analyze whether accuracy of TSR can be improved. Training Tiny YOLO with extended database does not improve its performance, and as a matter of fact, decreases it. Tiny YOLO have relatively simple architecture, which is very fast but more likely to overfit.

By careful inspection of detection results in Fig. 2, several interesting observations can be made. First of all, YOLOv3 does make very good detection and recognition of traffic signs with compact shape (rectangular, circle or hexagonal stop sign). In these cases, sometimes a double detection can occur. In the case of double detection, a two bounding boxes with different probability scores are obtained, which is clearly notable in the second, third and fourth row in Fig. 2. This happens because several grids include center of the sign, and high but different probability score calculated for both of them. In most cases double detection bounding boxes significantly overlaps, so adequate selection of non-maxima-suppression algorithm will remove undesirable detection's. Secondly, the worst performance YOLOv3 has with direction traffic signs (category VII), which differ from each other by an arrow direction and the titles (name of the place). All instances within this category have a similar visual appearance, but the big difference is between the left or right arrow direction. Also, the traffic signs usually have different titles, even in the same class.

An overview of per category mAP^{50} for both YOLOv3 and Tiny YOLO is given in the Table II. We can notice much worst performance for categories VI and VII, which

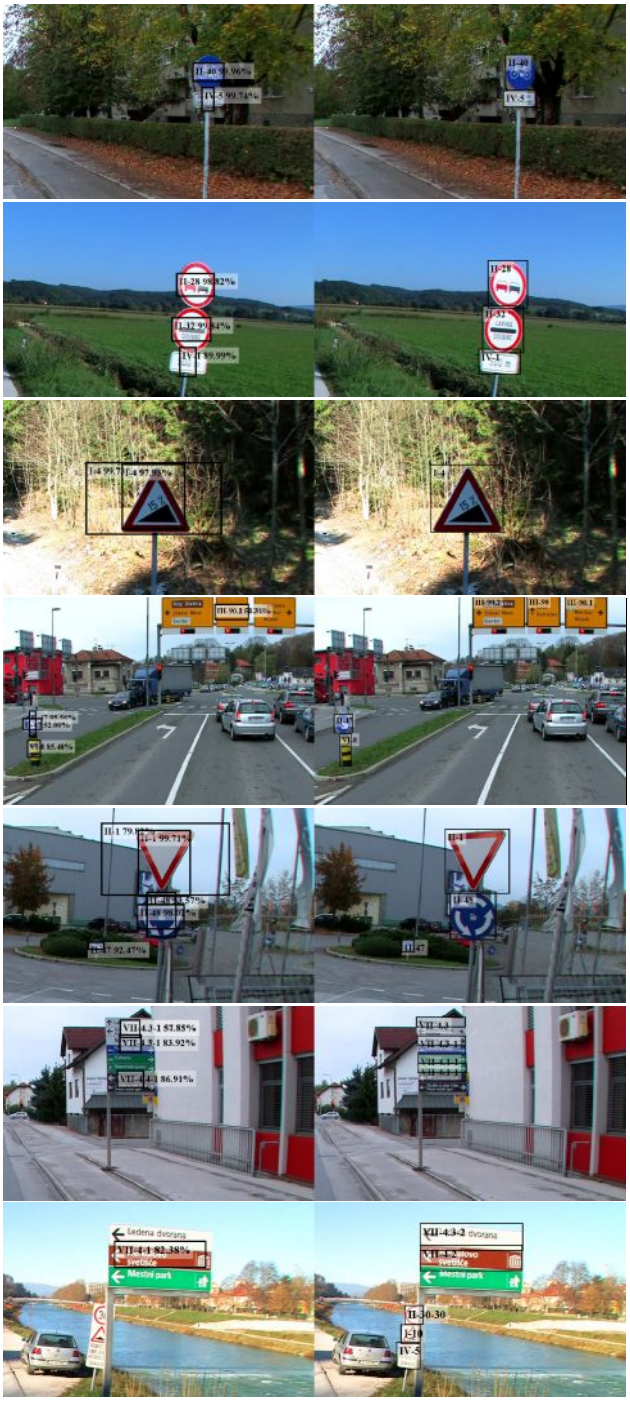

Fig. 2. Examples of traffic sign detection and recognition using YOLOv3. Images in the left column give traffic sign detections with corresponding probabilities, and images in the right column are ground truth annotations. (Best view in color.)

contain instances with different aspect ratio, poles with different stripes pattern and different arrow directions. Performance on categories with more compact-shape instances is much better.

IV. Conclusion

In this paper, we examined the possibility of real-time traffic sign detection and recognition using deep learning techniques. The results showed that real-time TSD/TSR is possible even for the HD images, with the mean averaged precision of 88%. As expected, the trade-off between and accuracy is inevitable and the result with faster version is less accurate. Also, traffic signs instances with compact aspect ratios are much easier to accurately detect and recognize.

References

[1] K. C. Wang, Z. Hou, and W. Gong, “Automated road sign inventory system based on stereo vision and tracking,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 25, no. 6, pp. 468–477.

[2] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, “Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition,” *Neural Networks*, vol. 32, pp. 323 – 332, 2012, selected Papers from IJCNN 2011.

[3] Y. Zhu, C. Zhang, D. Zhou, X. Wang, X. Bai, and W. Liu, “Traffic sign detection and recognition using fully convolutional network guided proposals,” *Neurocomputing*, vol. 214, pp. 758 – 766, 2016.

[4] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, “Traffic sign detection and classification in the wild,” in 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2110–2118.

[5] F. Zaklouta and B. Stanciulescu, “Real-time traffic-sign recognition using tree classifiers,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1507–1514, Dec 2012.

[6] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, “Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1484–1497, Dec 2012.

[7] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, “Detection of traffic signs in real-world images: The german traffic sign detection benchmark,” in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, Aug 2013, pp. 1–8.

[8] M. Haloi, “A novel plsa based traffic signs classification system,” *CoRR*, vol. abs/1503.06643, 2015. [Online]. Available: http://arxiv.org/abs/1503.06643

[9] P. Sermanet and Y. LeCun, “Traffic sign recognition with multi-scale convolutional networks,” in *The 2011 International Joint Conference on Neural Networks*, July 2011, pp. 2809–2813.

[10] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, “Multi-column deep neural network for traffic sign classification,” *Neural Networks*, vol. 32, pp. 333 – 338, 2012, selected Papers from IJCNN 2011.

[11] J. Jin, K. Fu, and C. Zhang, “Traffic sign recognition with hinge loss trained convolutional neural networks,” *IEEE Transactions on Intelligent Transportation Systems,* vol. 15, no. 5, pp. 1991–2000, Oct 2014.

[12] R. Timofte, K. Zimmermann, and L. V. Gool, “Multi-view traffic sign detection, recognition, and 3d localization,” in *2009 Workshop on Applications of Computer Vision (WACV)*, Dec 2009, pp. 1–8.

[13] S. Šegvic, K. Brkić, Z. Kalafatić, V. Stanisavljević, M. Ševrović, D. Budimir, and I. Dadić, “A computer vision assisted geoinformation inventory for traffic infrastructure,” in *13th International IEEE Conference on Intelligent Transportation Systems*, Sept 2010, pp. 66–73.

[14] G. Huang, Z. Liu, and K. Q. Weinberger, “Densely connected convolutional networks,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[15] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” *CoRR*, vol. abs/1506.02640, 2015. [Online]. Available: http://arxiv.org/abs/1506.02640

[16] J. Redmon and A. Farhadi, “YOLO9000: better, faster, stronger,” *CoRR*, vol. abs/1612.08242, 2016. [Online]. Available: http://arxiv.org/abs/1612.08242

[17] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv*, 2018.

[18] P. Uršič, D. Tabernik, R. Mandeljc, and D. Skočaj, “Towards large-scale traffic sign detection and recognition,” in *Proceedings of the 22nd Computer Vision Winter Workshop*, 2017.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, J. and A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge", *International Journal of Computer Vision*, 88(2), 303-338, 2010