Stereo Vision-Based Convolutional Networks for Object Detection in Driving Environments

Carlos Guindel(✉), David Martín, and José María Armingol

Intelligent Systems Laboratory
Universidad Carlos III de Madrid, Leganés, Spain
{cguindel, dmgomez, armingol}@ing.uc3m.es

Abstract. Deep learning has become the predominant paradigm in image recognition nowadays. Perception systems in vehicles can also benefit from the improved features provided by modern neural networks to increase the robustness of critical tasks such as obstacle avoidance. This work proposes a vision-based approach for on-road object detection which incorporates depth information from a stereo vision system within the framework of a state-of-art deep learning algorithm. Experiments performed on the KITTI benchmark show that the proposed approach results in significant improvements in the detection accuracy.

Keywords: Object detection · Stereo vision · Deep learning

1 Introduction

Detection of objects from a moving observer is an essential task for a large number of advanced driver assistance systems (ADAS) and virtually every autonomous car. As vehicles are meant to share the road with other users, each with its distinctive behavior, predictions about future traffic situations require an accurate identification of the objects in the surroundings.

While object detection in images is a classic problem in computer vision, traffic scenes are particularly complex due to the diversity of appearances, poses, and occlusions. Additionally, robustness to changes in illumination, weather, and other external factors is an implicit prerequisite for these applications. Challenges posed by driving environments have often been tackled making use of the additional information provided by stereo vision systems [1], which are composed of two nearly-identical cameras displaced horizontally from one another. This setup allows the extraction of depth information about the scene.

On the other hand, deep learning has become ubiquitous in almost every application involving image recognition in the past few years. Convolutional Neural Networks (CNNs) are currently the method of choice after they have demonstrated to be extremely useful in practical applications. Their success stems from their ability to learn hierarchical features which significantly outperform previous hand-crafted features for a variety of computer vision tasks.

In this work, we aim to enhance the performance of a state-of-art object detection framework, Faster R-CNN [2], by incorporating depth information from a stereo camera in a simple, straightforward way.

2 Related Work

The standard pipeline for object detection in images entails two main stages: extraction of regions of interest (ROIs) and classification of those proposals. A few years ago, the research interest was focused on hand-crafted features, e.g. HOG [3]. As the complexity of feature extraction schemes was tractable, it was usually possible to perform an exhaustive search over the image using a sliding window approach.

The introduction of CNNs led to a paradigm shift: today, features are learned in a supervised optimization process that makes use of large datasets. The complex hierarchical structures of CNNs involve longer computation times, and sliding-window approaches have become unfeasible due to the huge amount of regions to be classified. For this reason, as well as the large receptive fields featured by the conventional CNN architectures, extraction of ROIs in deep-learning-based object detection schemes remains a very active research area.

Girshick et al. [4] developed the R-CNN paradigm, where CNN features are computed for every candidate ROI and used in a further classification step. The method was further updated in [5] with the introduction of Fast R-CNN.

As a natural evolution, Faster R-CNN [6] extends the CNN approach to the ROI extraction stage, thus resulting in an end-to-end detection framework. The convolutional layers are applied over the image to extract features which are simultaneously used to propose candidate regions and to classify them. The former is performed by a Region Proposal Network (RPN), while the later is carried out with Fast R-CNN, which additionally provides a bounding box refinement. As a consequence, the most time-consuming task, i.e. the computation of the convolutional features, is performed only once.

Despite the impressive performance of Faster R-CNN in generic datasets, e.g. ILSVRC [7], achieved with only a fraction of the cost of more sophisticated models, hypothesis generation remains a substantial limiting factor in performance. As a matter of fact, a significant number of methods in the top positions of the challenging KITTI benchmark [8] are evolutions of the baseline Faster R-CNN approach specifically designed to overcome this limitation, such as the scale-dependent pooling introduced in [9], or the multi-scale CNN presented in [10].

3 Object Detection Approach

We aim to enhance the solid detection baseline provided by Faster R-CNN by leveraging the stereo depth information without significantly altering the original design. For that end, we adapt the setup of the network model to allow the processing of four-channel data structures containing the RGB color channels of the left image and, additionally, a scaled disparity map. Our approach is summarized in Fig. 1.

The disparity map is a data structure which encodes the deviation in horizontal coordinates, d, of corresponding points in both images belonging to the

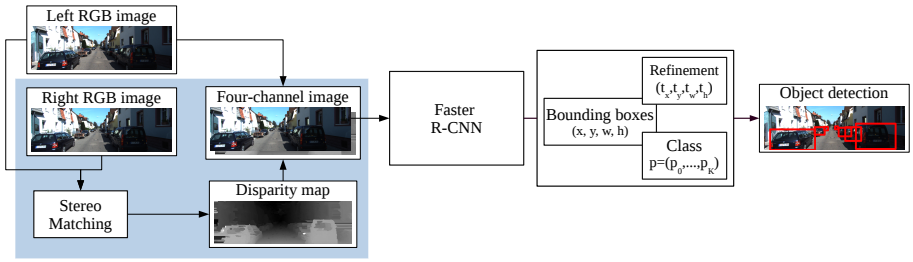

Fig. 1. Proposed object detection method. Our contribution is highlighted in blue.

stereo pair. Thus, the value of each pixel in our *fourth channel*, $s \cdot d$, is inversely proportional to the scene depth at that location, Z, following the relation:

$$s \cdot d = \frac{f \cdot B}{Z} \tag{1}$$

where f is the focal length and B the baseline of the binocular pair. Since these values are determined for a particular stereo system, the disparity value is indeed inversely proportional to the actual depth.

The reasoning behind our approach is that region proposal can take advantage of the geometrical information provided by the disparity estimation to segment the foreground objects from the background, thus overcoming the most severe shortcoming of the CNN method. Using disparity values straightforwardly, instead of actual depth values, is expected to benefit the segmentation of objects at closer distances due to the inverse relationship linking both magnitudes.

In summary, our design is intended to preserve the end-to-end nature of the Faster R-CNN detection method while enhancing the performance of the classification, especially for objects represented with a limited number of pixels.

3.1 Parameter Tuning

Before considering the influence of the depth channel in the CNN architecture, we optimized the performance of the baseline Faster R-CNN by tuning its hyper-parameters according to the specific requirements of driving environments. The modifications are targeted to the KITTI dataset [8], and include:

1. **Training samples selection.** Samples used in the training procedure are chosen so that their IoU overlap with any ground-truth *DontCare* label, corresponding to distant or unclear objects, is below a certain threshold: 25% for the Fast R-CNN module and 15% for the RPN. On the other hand, only samples eligible to be included in the 'hard' difficulty level are used.
2. **Scale.** Faster R-CNN has been shown [11] to be highly sensitive to the size of the input images. We have found that scaling the original images (with resolutions around 1242×375) to 500 pixels in height, both for training and evaluation, offers a good trade-off between accuracy and computation time.

3. **RPN anchors.** Proposals from the RPN are parametrized relative to fixed boxes called *anchors*. The design of the RPN is intended to handle scales and aspect ratios different than those of the anchors; however, using anchors of multiple sizes has been proven as an effective solution, so the a-priori knowledge about the objects in the environment can be used to further improve the detection accuracy. We use three scales and three aspect ratios for the RPN anchors, as in the original Faster R-CNN; but the values have been modified to fit the typical traffic participants, according to Table 1.

Table 1. Modification in the settings of RPN anchors.

	Original	Proposed
Scales	$\{128^2, 256^2, 512^2\}$	$\{80^2, 112^2, 144^2\}$
Aspect ratios	{2:1, 1:1, 1:2}	{5:2, 5:4, 2:5}

3.2 Stereo Depth Information

Different alternatives can be adopted to estimate the disparity map from the images of the stereo pair. Henceforth, the following methods are considered:

1. The classical Semiglobal Matching algorithm [12] in its OpenCV implementation [13]; i.e. using block matching and the Birchfield-Tomasi metric.
2. A state-of-art CNN-based algorithm, DispNet [14], currently ranked 9th in the KITTI stereo leaderboard among the published methods[1] and with a reported runtime of 60 ms.

The density of the SGM disparity map is around 90% due to the existence of unmatched pixels. As these *undefined* values could prevent the gradient descent training to converge, we perform a background interpolation to fill the holes, so every pixel (u_0, v) with a undetermined value in the disparity map, $\nexists d(u_0, v)$, is given a value according to:

$$\hat{d}(u_0, v) = \min(d(u_0^-, v), d(u_0^+, v)) \tag{2}$$

where $d(u_0^-, v)$ and $d(u_0^+, v)$ are the disparities of the contiguous *defined* pixels in the same row. DispNet, on the other hand, provides a 100% dense disparity map.

As mentioned above, values in the disparity map are scaled before entering the CNN, according to the s factor in Eq. 1. This is actually a normalization of the disparity values between 0 and $255/s$. Note that the scaling operation must be performed with saturation to prevent overflow. We chose $s = 4$ in order

[1] `http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo`

to obtain values close to the pixels in the color channels; this means that only disparities originally in the range between 0 and 64 are distinguishable in the resulting map. Given the parameters of the KITTI stereo system, that clipping corresponds to depths from 6m to the infinite, which is reasonably tailored to the field of view of the camera. Fig. 2 depicts an example of the resulting fourth channel (already normalized) for each of the two employed stereo matching approaches.

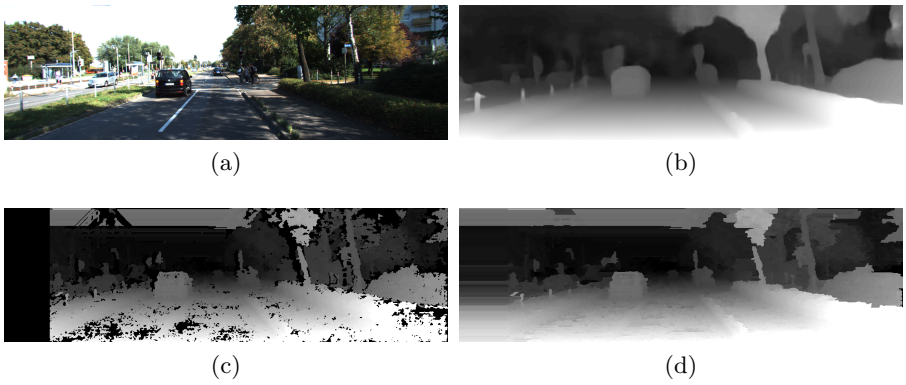

Fig. 2. Example of normalized disparity maps for a frame in the KITTI dataset (a), obtained with the two selected stereo matching approaches: DispNet [14] (b) and SGM [12] with interpolation (d), computed from the original SGM (c).

CNN architectures typically used in image recognition can be applied to our approach with minimal changes: only the filters in the first convolutional layer have to be adapted to accept a four-channel input.

A common practice in training CNN models is to initialize the weights in the convolutional layers using values trained in larger datasets, such as the ILSVRC [7], with the hope that the learned features may still be useful for related applications. As the filters that we use in the first convolutional layer are different from the existing pre-trained models, we initialize the weights in the fourth channel as the mean value of the same weight in the filters corresponding to the preexisting color channels. This approach, which avoids the need to retrain the models from scratch, is based on the assumption that discontinuities in depth are related to discontinuities in intensity. Additionally, we let weights in all the convolutional layers, including the shallower ones, be modified during training to fit the new nature of the data.

4 Results

We compare our approach with the baseline Faster R-CNN to investigate the improvement introduced by the stereo information. We use the already mentioned

KITTI object detection benchmark [8] for evaluation. Since the test ground-truth labels are not publicly available, we use the train/validation split by [15] to ensure that images from the same sequence do not exist in both training and validation sets. Following the standard KITTI setup, we use the Average Precision (AP) metric to evaluate the performance of the object detection pipeline and require IoU overlaps of 70%, for cars, and 50%, for pedestrian and cyclists.

We employ the VGG16 architecture [16], with the minimal changes discussed in Section 3.2. The approximate joint training from [2] is adopted. For every method, training has been performed for 50k iterations with a learning rate of 0.001 and then for 30k iterations with 0.0001. The seven distinct categories in the KITTI dataset are considered; however, only *Car*, *Pedestrian* and *Cyclist* classes are evaluated because of the low number of samples in the remaining categories. The number of RPN proposals is limited to 300; additionally, a non-maximum suppression (NMS) is performed. Results are presented in Table 2 for every category and level of difficulty.

Table 2. Detection AP (%) obtained on the KITTI validation set.

Input	Easy	Moderate	Hard
	Car		
RGB	88.76	77.01	60.81
RGB+SGM	**89.39**	**77.99**	**66.84**
RGB+DispNet	88.82	77.29	66.56
	Pedestrian		
RGB	85.97	68.71	61.41
RGB+SGM	87.39	69.16	63.62
RGB+DispNet	**87.70**	**69.73**	**64.47**
	Cyclist		
RGB	65.22	53.67	50.37
RGB+SGM	64.07	52.25	49.68
RGB+DispNet	**66.51**	**55.77**	**52.26**

Detection using the disparity information surpasses the bare RGB approach in almost all cases, with the notable exception of SGM for cyclists. On the other hand, DispNet outperforms the SGM estimation for pedestrians, while SGM shows better results for cars. The improvement introduced by the disparity information is especially noticeable in 'hard' samples, as shown in the summary tabulated in Table 3.

The average running time per image of the detection stage is 116 ms using a NVIDIA Titan Xp and Caffe [17]. For preliminary results with fixed weights in the first two convolutional layers during training, please refer to the Extended Abstract of this paper [18].

Table 3. Summary of mAP (%) obtained on the KITTI validation set, expressed as the difference in percentage points from the baseline RGB approach.

Input	Easy	Moderate	Hard
RGB	79.98	66.46	57.53
RGB+SGM	+0.30	+0.01	+2.52
RGB+DispNet	**+1.03**	**+1.14**	**+3.57**

5 Conclusion and Future Work

We have presented an approach to exploit the spatial information provided by a stereo vision system in order to enhance a well-established object detection method based on Convolutional Neural Networks. Our proposal is particularly suitable for automotive applications, where stereo cameras have frequently been employed to deal with the complexity of the environments without significantly altering the features of the vehicle.

Results have proven the potential of stereo information to enhance the convolutional features produced by the network, leading thus to a significant enhancement of the detection performance. The improvement is especially notable when detecting Vulnerable Road Users (VRU), namely pedestrians and cyclists, frequently identified as the most problematic categories in image recognition.

Further steps might focus on the architecture of the network, adopting either modern architectures, e.g. ResNets, or ad-hoc designs intended to exploit the information extraction from the disparity map. Additionally, some of the developments recently introduced in the literature to overcome the fixed size of the receptive field could be adopted.

This work is intended to be the first step towards a full scene understanding system in our IVVI 2.0 intelligent vehicle [19], an experimental platform for driving assistance systems. This application, along with other critical perception modules, will enable inference about complex traffic situations.

Acknowledgments Research supported by the Spanish Government through the CICYT projects (TRA2015-63708-R and TRA2016-78886-C3-1-R), and the Comunidad de Madrid through SEGVAUTO-TRIES (S2013/MIT-2713). The Titan Xp used for this research was donated by the NVIDIA Corporation.

References

1. Bernini, N., Bertozzi, M., Castangia, L., Patander, M., Sabbatelli, M.: Real-time Obstacle Detection using Stereo Vision for Autonomous Ground Vehicles: A Survey. In: IEEE International Conference on Intelligent Transportation Systems (ITSC). (2014) 873–878
2. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6) (2016) 1137–1149

3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 1. (2005) 886–893
4. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) 580–587
5. Girshick, R.: Fast R-CNN. In: Proc. IEEE International Conference on Computer Vision (ICCV). (2015) 1440–1448
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Advances in Neural Information Processing Systems (NIPS). (2015)
7. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3) (2015) 211–252
8. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 3354–3361
9. Yang, F., Choi, W., Lin, Y.: Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 2129–2137
10. Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N.: A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In: Computer Vision - ECCV 2016. Lecture Notes in Computer Science, vol 9908. (2016) 354–370
11. Fan, Q., Brown, L., Smith, J.: A Closer Look at Faster R-CNN for Vehicle Detection. In: Proc. IEEE Intelligent Vehicles Symposium (IV). (2016) 124–129
12. Hirschmüller, H.: Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(2) (2008) 328–341
13. Kaehler, A., Bradski, G.: Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library. O'Reilly Media, Inc. (2016)
14. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 4040–4048
15. Chen, X., Zhu, Y.: 3D Object Proposals for Accurate Object Class Detection. In: Proc. Advances in Neural Information Processing Systems (NIPS). (2015) 424–432
16. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR **abs/1409.1** (2014)
17. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding. In: Proc. ACM International Conference on Multimedia. (2014) 675–678
18. Guindel, C., Martín, D., Armingol, J.M.: Stereo Vision-Based Convolutional Networks for Object Detection in Driving Environments. In: EUROCAST 2017 - Extended Abstracts. (2017) 288–289
19. Martín, D., García, F., Musleh, B., Olmeda, D., Peláez, G.A., Marín, P., Ponz, A., Rodríguez Garavito, C.H., Al-Kaff, A., de la Escalera, A., Armingol, J.M.: IVVI 2.0: An Intelligent Vehicle based on Computational Perception. Expert Systems with Applications **41**(17) (dec 2014) 7927–7944