Object Detection Based on Deep Learning and Context Information

Paulin Pekezou Fouopi, Gurucharan Srinivas, Sascha Knake-Langhorst, and Frank Köster

German Aerospace Center, Institute of Transportation Systems, Braunschweig, Germany

{paulin.pekezoufouopi,gurucharan.srinivas,sascha.knake-langhorst,frank.koester}@dlr.de

Abstract. In order to avoid collision with other traffic participants, automated vehicles need to understand the traffic scene. Object detection, as part of scene understanding, remains a challenging task mostly due to the highly variable object appearance. In this work, we propose a combination of convolutional neural networks and context information to improve object detection. To accomplish that, context information and deep learning architectures, which are relevant for object detection, are chosen. Different approaches for integrating context information and convolutional neural networks are discussed. An ensemble system is proposed, trained, and evaluated on real traffic data.

Keywords: Object Detection, Convolutional Neural Networks, Context Information, Bayesian Models

1 Introduction

Automated driving is one of the most important research topics in automotive area. In recent years, many projects like PROMETHEUS, the DARPA Grand/Urban challenge, and CityMobil as well as different research groups and institutions have addressed this topic with promising results. To plan a collision free trajectory, automated driving vehicles must be able to detect objects. Although many solutions are available in the literature, this remains a challenging task due to huge variation in object appearance and scene complexity. Object appearance can change according to occlusion, noise, variation in pose and illumination [1], and background clutter. Convolutional Neural Networks (CNN) show the best classification results, but still have some classification errors because they are mostly appearance-based classifiers. Context information can be used to improve object detection [1]. In this paper we propose an object detection system, which uses the advantages of CNN and context-based classifiers. We discuss different approaches for combining both classifiers. The proposed system is trained and evaluated on real traffic data.

The next sections of this work are divided as follows: in section 2, we present the state of the art. The proposed system as well as training and evaluation results are discussed in section 3. In section 4, we conclude and give an outlook on future work.

2 Related Work

Object detection consists of localizing object instances (hypotheses generation) in an image and classifying those into semantic classes (hypotheses classification). Hypotheses are generated using features like symmetry, aspect ratio, expected position, color, and motion. Hypotheses classification methods can be separated into shape- and feature-based approaches. In this work we focus on the second one.

Feature-based approaches first transform hypotheses into features and classify them. Features can be generated manually or learned directly from the data using e.g. Deep Learning (DL). Manually generated features like Histogram of Oriented Gradients and Deformable Parts Model [2] are used with Shallow Learning (SL) classifiers like Support Vector Machine for vehicle and pedestrian classification. While these SL-classifiers show promising results, they suffer from human errors made during the feature engineering task. DL approaches solve this problem by learning the specific features inherently from large training data set. Since 2012, many DL-classifiers like Faster R-CNN and Yolo outperform SL-classifiers for object detection, but suffer from wrong detections mostly due to the appearance variation drawback depicted above.

In [1, 3] spatial (interposition, support, and position), semantic (co-occurrence), and scale (familiar size) context information between objects, scenes und situations were combined with SL-classifiers to improve object detection. It is difficult to explicitly model the contextual dependencies described above into CNN because CNN just reason about spatial dependencies between object parts. The simplest solution is to integrate context information as pre- or post-processing step. Chu et al. [4] used an ensemble system, which combined the Faster R-CNN, local and global context for object detection. Some efforts to integrate context information directly into the CNN were shown in [5] (time constraint) and [6] (global image-level and local super-pixel context). Liang et al. [7] argued that Recurrent CNN (RCNN) were more suitable for integrating contextual relations, but RCNN can just reason about spatial dependencies between objects and their parts. Contextual dependencies on object and scene levels were still missed and will be address in this work.

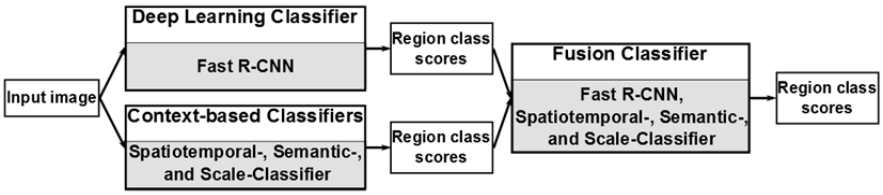

Fig. 1. Overview of the system for integrating DL and context-based classifiers

3 "Our Approach" with Current Results

In this work, we focus on the integration of DL and context-based classifiers using an ensemble system (see **Fig. 1**). We follow the idea proposed in [4], but use different context information and graphical model. As DL-classifier, we choose the pre-trained Fast R-CNN [8]. The semantic (se_{cf}), spatiotemporal (st_{cf}) and scale (sc_{cf}) context proposed in [1] are used for generating context-based features. The context-based classifiers estimate the conditional class probability $p(C|X_{\mathrm{cf}})$ of an object hypothesis given the context-based feature $X_{\mathrm{cf}} \in \{se_{\mathrm{cf}}, st_{\mathrm{cf}}, sc_{\mathrm{cf}}\}$ using the naïve Bayes classifier

$$p(C|X_{\mathrm{cf}}) = \frac{p(X_{\mathrm{cf}}|C)p(C)}{\int_C p(X_{\mathrm{cf}}|C)p(C)}, \tag{1}$$

where $p(X_{\mathrm{cf}}|C)$ is the likelihood function. $C \in \{ped., non\_ped.\}$ is the semantic class set and $p(C)$ the prior class probability. The fusion classifier combines the Fast R-CNN and the context-based classifiers scores S_{f_rcnn} and $S_{cb\_c} = (S_{se\_c}, S_{st\_c}, S_{sc\_c})$ using a Bayesian network and the assumption that the scores are conditionally independent given C. The conditional class probability is

$$p(C|S_{cb\_c}, S_{f\_rcnn}) = \frac{p(S_{cb\_c}|C)p(S_{f\_rcnn}|C)p(C)}{\int_C p(S_{cb\_c}|C)p(S_{f\_rcnn}|C)p(C)}. \tag{2}$$

$p(S_{cb\_c}|C)$, and $p(S_{f\_rcnn}|C)$ are the likelihood functions. S_{se_c}, S_{st_c}, and S_{sc_c} are the semantic, spatiotemporal and scale context-based classifiers scores.

For evaluating the proposed system, we used the Caltech Pedestrian Data Set (CPDS) [9]. Only the aspect ratio $a_r = w/h$ was used as context-based feature, since it belonged to the scale context proposed in [1] and the CPDS didn't contain depth information. h and w were the height and the width of a given bounding box. The likelihood functions $p(X_{\mathrm{cf}}|C)$, $p(S_{cb\_c}|C)$, and $p(S_{f\_rcnn}|C)$ were modeled as Gaussian distributions and the Maximum Likelihood Estimator (MLE) were used to estimate theirs parameters. The prior probability $p(C)$ was the ratio of pedestrians respectively non-pedestrians present in the training dataset. **Fig. 2** presents from left to right the ground truth as well as the Fast R-CNN, the aspect ratio-based classifier (A_R-classifier), and the fusion classifier results with scores greater than 0.5. We observed that the Fast R-CNN detected the most of pedestrians. Just a few objects were missed probably because of the low resolution and occlusion. Although the A_R-classifier had many false positive, the fusion classifier improved the Fast R-CNN and A_R-classifier detecting more pedestrians. The fusion classifier false positive could be explained by the fact that aspect ratio was not powerful enough to model the context.

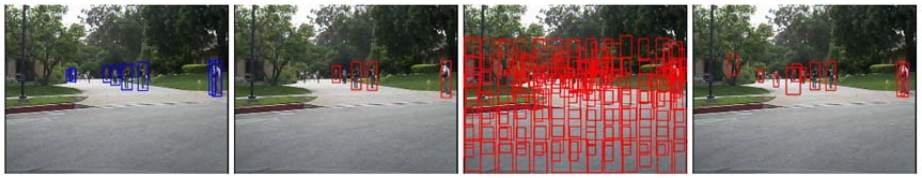

Fig. 2. Detection results on CPDS ([9]). See text for more information.

4 Conclusion and Future Work

In this work, we addressed the problem of integrating context information and DL architectures into a system for object detection. A fusion system combining DL and context-based classifiers was proposed. We modeled the context-based classifiers using the naïve Bayes method. The DL and the context-based classifiers scores were fused using a Bayes model. For training and evaluating our system, we used the DL-classifier called Fast R-CNN. The context-based features were generated using aspect ratio. The Likelihood functions parameters were learned with the MLE on the CPDS dataset. First results on real data revealed that the proposed system was able to improve the detection in some cases, but also had some false positive. Integrating more context information may compensate this effect.
In our future work, we will integrate more context information (e.g. explicitly reasoning about occlusion) and evaluate the system on large data set. The problem of integrating context directly into the DL architecture will be addressed. Another key aspect will be to investigate the possibility of learning context information directly from the data without explicit modelling.

References

References
1. Galleguillos, C., Belongie, S.: Context Based Object Categorization: A Critical Survey. Comput. Vis. Image Underst. 114, 712–722 (2010)
2. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part-Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645 (2010)
3. Biederman, I., Mezzanotte, R.J., Rabinowitz, J.C.: Scene perception. Detecting and judging objects undergoing relational violations. Cognitive Psychology 14, 143–177 (1982)
4. Chu, W., Cai, D.: Deep Feature Based Contextual Model for Object Detection. CoRR abs/1604.04048 (2016)
5. Kang, K., Ouyang, W., Li, H., Wang, X.: Object Detection from Video Tubelets with Convolutional Neural Networks. CoRR abs/1604.04053 (2016)
6. Liang, X., Xu, C., Shen, X., Yang, J., Tang, J., Lin, L., Yan, S.: Human Parsing with Contextualized Convolutional Neural Network. IEEE Transactions on Pattern Analysis and Machine Intelligence PP (2016)
7. Liang, M., Hu, X. (eds.): Recurrent convolutional neural network for object recognition. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
8. Girshick, R.B.: Fast R-CNN. CoRR abs/1504.08083 (2015)
9. Caltech Pedestrian Detection Benchmark, http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/