Co-training of context models for real-time object detection

Alexander Gepperth

To cite this version:

Alexander Gepperth. Co-training of context models for real-time object detection. IEEE Symposium on Intelligent Vehicles, Jun 2012, Madrid, Spain. hal-00763676

HAL Id: hal-00763676

https://hal.inria.fr/hal-00763676

Submitted on 11 Dec 2012

Co-training of context models for real-time vehicle detection

Alexander R.T. Gepperth[1,1]

ENSTA ParisTech, 32 Blvd Victor, 70539 Paris Cedex 15

Abstract

We describe a simple way to reduce the amount of required training data in context-based models of real-time object detection. We demonstrate the feasibility of our approach in a very challenging vehicle detection scenario comprising multiple weather, environment and light conditions such as rain, snow and darkness (night). The investigation is based on a real-time detection system effectively composed of two trainable components: an exhaustive multiscale object detector ("signal-driven detection"), as well as a module for generating object-specific visual attention ("context models") controlling the signal-driven detection process. Both parts of the system require a significant amount of ground-truth data which need to be generated by human annotation in a time-consuming and costly process.

Assuming sufficient training examples for signal-based detection, we demonstrate that a co-training step can eliminate the need for separate ground-truth data to train context models. This is achieved by directly training context models with the results of signal-driven detection. We show that this process is feasible for different qualities of signal-driven detection, and maintains the performance gains from context models.

As it is by now widely accepted that signal-driven object detection can be significantly improved by context models, our method allows to train strongly improved detection systems without additional labor, and above all, cost.

1. Introduction

Our experience with cluttered and uncontrolled traffic environments[5, 7, 6, 4] suggests that purely appearance-based (i.e. based on local pixel patterns) object detection suffers from ambiguities. We claim that object-specific models relating appearance-based visual information to non-local and non-visual information must be taken into account to achieve the required disambiguation. We shall denote such models "context models".

Messages of the article. In a previous work[7], we have shown by benchmarks how context models can be acquired and exploited for real-time vehicle detection, and that their application is both simple and beneficial in terms of detection performance. This investigation is based on the vehicle detection system of [7] whose overall organization is depicted in Fig. 1. The goal of the article is to elabo-

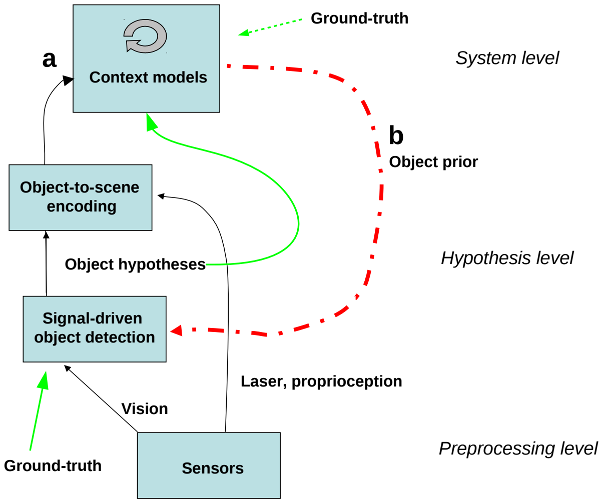

Figure 1: Structural overview of the system described in [7]. The red arrow indicates the reverse propagation of object priors that are derived from trained context models. The solid green arrow shows where training data is supplied to the system, whereas the dotted arrow indicates the additional requirement for ground-truth data when not co-training context models.

Email address: alexander.gepperth@ensta-paristech.fr (Alexander R.T. Gepperth)

rate upon the fact that context models can be successfully trained even in the absence of error-free ground-truth data (an approach usually termed *co-training*, see, e.g., [15]), and to shed light on the conditions under which such a process may be expected to work.

Evaluation data. All experiments are conducted using the publicly available HRI RoadTraffic dataset[7], containing five extended, annotated video sequences of inner-city driving, comprising a large variation of environmental and image processing conditions[1].

Approach. Given an object hypothesis produced by signal-driven detection, context models (as used in [7]) are essentially classifiers that estimate the identity of the hypothesis independently of its visual appearance, see Fig. 8. Input for this estimation are *object-to-scene properties* which are computed from scene information and the location of the object hypothesis. Typical object-to-scene properties include, e.g., the height over the road plane, the distance to the obstacle-free road area, or simply the position in the 2D camera image. Therefore, context model training requires object object identity to be known for each object hypothesis, which is usually achieved by using manually created *ground-truth data*. The basic idea of context model co-training is to replace ground-truth data with the object identity estimate from signal-driven object detection, which is available ”for free”. In general, this identity estimate will be error-prone; however, if it is correct ”on average”, it is intuitive that successful training of context models could be feasible. To assess this, the experiments of this article are conducted in two conditions which differ in the way the signal-driven detection algorithm of [22] is trained. In the *default condition*, training is done using ground-truth vehicle data taken from the HRI RoadTraffic dataset[7]. In the *impaired condition*, data of inferior quality generated by a laser sensor is employed, leading to reduced detection performance. By co-training context models and and performing a subsequent performance evaluation in both conditions, we determine whether co-training of context models becomes infeasible with decreasing detection performance.

Related work. In this article, we use a co-training strategy[2] where one classifier trains another. This approach has been applied to challenging object detection scenarios, e.g., person detection[16], vehicle detection for traffic surveillance[1] or face detection[11, 20]. Similar to [1], the application target of our work is vehicle detection, although we assume a moving instead of a static platform. A further similar point is the use of a primitive vehicle detector (laser in our case, audio in [1]). Our context model approach couples the global spatial scene layout (”context”) to local object detection strategies. Such a coupling of object detection and contextual information has been used previously to improve object detection in a variety of scenarios ranging from controlled indoor scenes to realistic traffic environments[7, 13, 8, 3, 21, 19]. In [13], it is demonstrated that the ”gist”, i.e., a low-dimensional description of a scene, can be used to infer the locations of vehicles, signs and pedestrians in traffic scenes by statistical models constructed from training examples. The concept of gist is taken further in [8] where a generic probabilistic model of 3D scene layout is proposed that can be queried for likely image locations of, e.g., vehicles or pedestrians in order to inform an exhaustive local object detector. An approach that is somewhat related to our context models is presented in [18] where information about road area is used to infer likely locations of pedestrians after a supervised training process. Scene geometry estimation is used to compute prior distributions for vehicles and pedestrians in [10], although the probabilistic models are designed not learned. The presented work is based on the object detection architecture introduced in [7], focusing on the issue of bootstrapping. Going beyond [7], this article investigates the robustness of the bootstrapping process to less-than-ideal training data, and thus opens the possibility to train the combined detector-context system to without ground-truth data at all.

2. Outline

In Sec. 3, we will briefly describe the object-to-scene properties that are computed for each object hypothesis, and which serve as input to context models. Subsequently, we will sketch the working of laser-based object detection used to obtain training data without going into details due to space constraints. Afterwards, we will give an overview of

[1] For obtaining the annotated image data, please write an email to alexander@gepperth.net or hri-road-traffic@honda-ri.de

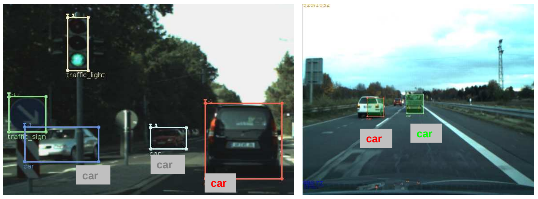

Figure 2: Visualization of the two different types of training data used to train signal-driven object detection. In the *default condition*, human-annotated ground-truth data is used for training (left), whereas the *impaired condition* uses data derived from laser sensors (right).

the training of context models and indicate how co-training is integrated into it. As we use the system of [7], we refer the interested reader to this reference for a more detailed description of this system. In Sec. 4, we will describe the conducted experiments, from which we will draw conclusions in Sec. 5.

3. Methods

3.1. Generation of training data for classifier

In contrast to the *standard condition* where data to train the signal-driven object detection are taken from the HRI RoadTraffic dataset, the *impaired condition* obtains training data by means of two laser scanners of the model ibeo LUX (see [9]) on an additional stream of highway driving which is not part of HRI RoadTraffic. Using elementary image processing and tracking techniques, we identify self-moving segments in the laser signal which we assume to be cars, and which we transform into image coordinates by means of the *camera transform* [17]. In this way, signal-driven detection can be trained with 3000 positive and 10000 random negative examples. Please see Fig. 2 for a visualization of the two types of training data, and Fig. 3 for an idea of laser processing (not described here).

3.2. Free-area computation

The so-called *free area* is defined as the obstacle-free space in front of the car that is visually similar to a road. This quantity carries significant semantic information. Since it is, by construction, bounded by all obstacles that the car might collide with, many relevant obstacles are close to the boundaries of the free area. The object-to-scene quantity of interest is therefore the image-based *distance* between an object hypothesis and the free area. Please see [12] for details of calculating the free-area and Fig. 4 for examples of distance-to-free-area measurements.

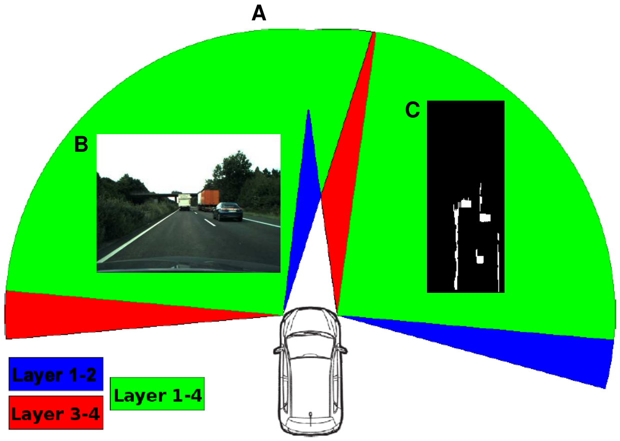

Figure 3: **A)** Setup of ibeo LUX laser sensors in the prototype vehicle used to create all recordings. To compensate for vehicle roll, each laser sensor measures distances in 4 layers which cover a vertical angle of 3.2 degrees (0.8 degrees per layer). Due to technical reasons, the lasers are built into the car in a slightly asymmetric fashion which results in different covered areas on the left and right side of the car. Since both devices can only use half of their layers at the borders of their angular range, the area in front of the car cannot be covered by 4 layers in both laser sensors. The effective angular resolution of both laser devices is 0.25 degrees. **B), C)** Camera image and clustering image obtained by the preprocessing of laser sensor results.

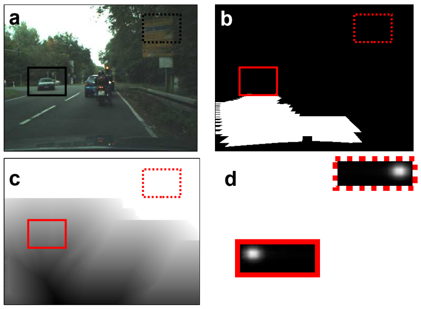

Figure 4: Distance to free area computation. **a)** original image with two object hypotheses **b)** computed free area **c)** pixelwise distance-to-free area map. Each pixel value in the map is determined by that pixel's minimal distance to a free-area pixel where an upper limit $d_{\max}$ is imposed for efficiency **d)** Visualization of object-to-scene feature value between 0.0 (left) and 1.0 (right).

3.3. Distance and elevation computations

We employ dense correspondence-based stereo processing for measuring the object-to-scene quan-

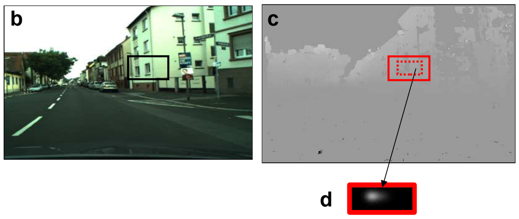

Figure 5: Elevation processing. **a)** video image with object hypothesis **b)** dense elevation map **c)** Visualization of object-to-scene feature value (0.0 = left, 1.0 = right).

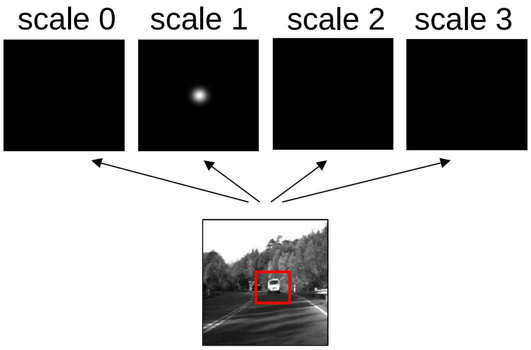

Figure 6: Size-dependent encoding of hypothesis position. Hypothesis size determines at which ”pyramid” level the position of the hypothesis is encoded.

tities of hypothesis distance and height in car-centered coordinates. The quantity that really carries semantic information is the *elevation* of object hypotheses, i.e., their height over the detected road surface. Please see Fig. 5 for an example of computing elevation, and [7] for further details.

3.4. Position and size related analysis

Two important although almost trivial object-to-scene quantities are the hypothesis position and size in the camera image. Even though the image position of objects changes, for example, during turning maneuvers (similar examples can be mentioned for image-based size), we found that these quantities can nevertheless provide useful hints about object identity. Therefore, they are encoded at the hypothesis level of our system as shown in Fig. 6.

3.5. Signal-driven object detection

Signal-based object detection generates object hypotheses in two successive steps. As a first step, a hierarchical feed-forward network is applied to the camera image in the manner of a convolutional network[22]. This produces a pyramid of K retinotopic confidence, or, if we wish to stay in the language of Bayesian inference, object likelihood maps. Each pixel of such a map indicates the presence of a specific view of an object (in our case: back-views of cars) at a specific scale, see Fig. 7. A list of object hypotheses is subsequently generated from the object likelihood maps by a competitive selection process described in [7].

3.6. Training and co-training of context models

Context models are trained in a supervised fashion using simple logistic regression models[7] as shown in Fig. 8. Instead of requiring additional ground-truth data, the supervision signal is derived from object identity as computed from signal-driven detection. For this purpose, the detection threshold is set very low such that a great number of hypotheses if produced. We consider each detection whose detection likelihood exceeds a threshold of 0.4 to be a vehicle for the purposes of training context models.

The trained context models can then be *inverted* to produce a ”object prior map” which is combined with the object likelihood maps obtained from the signal-driven detection, to yield a ”object posterior map” indicating the belief of vehicle detections at various locations. The effect of applying the object prior map in this way is demonstrated in Fig. 7.

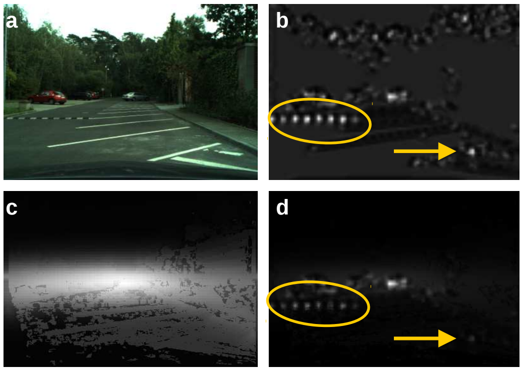

Figure 7: Typical effects of object priors on classifier output. **a)** Sample input image. **b)** object likelihood map of classifier at scale 5. Note the strong (but incorrect) maxima indicated by the ellipse and the arrow. **c)** Object prior map derived from context models at scale 5. **d)** Object posterior map. Note that the local maxima indicated by the arrow and the ellipse have been merely attenuated; especially the maximum indicated by the arrow may still be selected since there are no competing maxima nearby. In contrast, local maxima close to the upper border of the image have been eliminated.

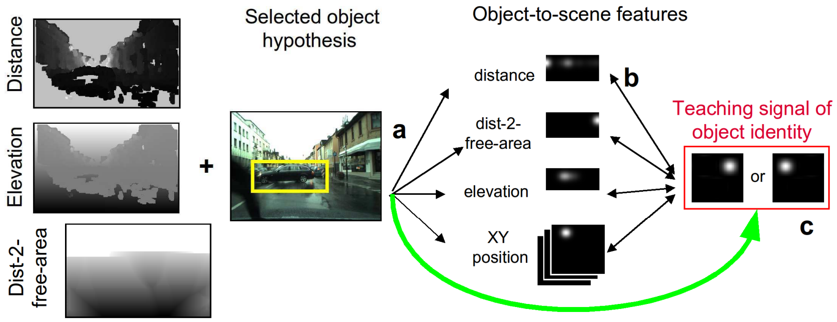

Figure 8: Training of context models. **a)** Hypothesis selected by signal-driven object detection. Object identity is used for co-training context models (green arrow) whereas hypothesis position is used for computing object-to-scene properties. **b)** Training the mapping between object-to-scene properties and object identity. **c)** Teaching signal for context models, derived from signal-driven object detection.

ID	weather	daytime	images	ann. images
I	overcast,dry	afternoon	9843	957
II	low sun, dry	late afternoon	22600	949
III	heavy rain	afternoon	6725	643
IV	dry	midnight	6826	464
V	after heavy snow	afternoon	16551	867

Table 1: Left: Details about the used video streams. Right: example images.

3.7. Experimental setup

The HRI RoadTraffic dataset which we use for evaluation contains five distinct color video streams (denoted I-V) together with laser range finder data for free area detection. All videos are around 15 minutes in length, and were taken during test drives along a fixed route covering mainly inner-city areas, along with short times of highway driving. Please see Tab. 1 for details and a visual impression.

For performance assessment, we compute receiver-operator characteristics (ROCs) combining common [14, 7] evaluation measures, namely the *false negative rate (fnr)* and *false positives per image (fppi)*.

4. Experiments and Results

For all experiments, the training of context models is performed using a procedure called *blocking*: we group the stream of hypotheses into intervals corresponding to 30s of real time and apply context model training only for odd-numbered groups. The even-numbered groups are used for performance evaluation, which ensures that training and evaluation data are always strictly non-overlapping.

For each experiment, we run the system twice, assuming that signal-driven detection has already been trained. In the first run, context models are trained, and therefore the generation of (untrained) object priors is disabled. The detection threshold of signal-driven detection is set to 0.0, and context models are trained with a ”vehicle” signal for each hypothesis whose likelihood exceeds 0.4. In the second run, learning is switched off, and the effect of trained context models on detection is evaluated on streams I-V by computing ROC-like plots (see Sec. 3.7).

Two experiments are conducted, one for the default condition and one for the impaired condition using laser-generated training data for signal-driven object detection (see Sec. 1). For the impaired condition, we only show results on stream III for space limitations, and since results are redundant. Fig. 9 and Fig. 10 visualize the measured system performances in both conditions.

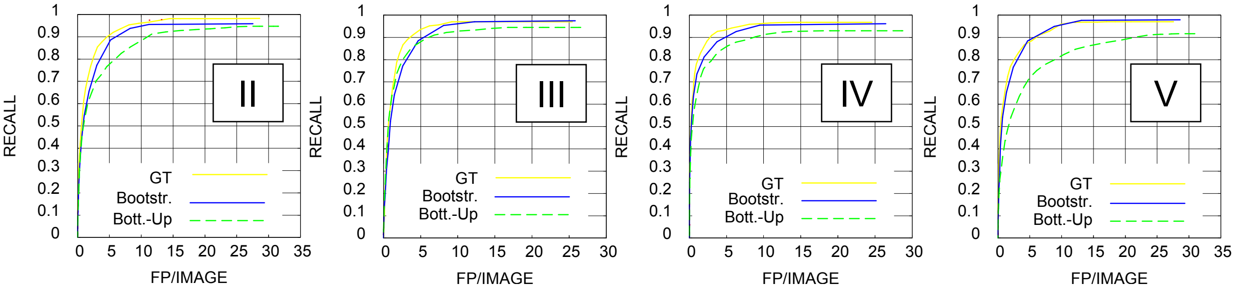

Figure 9: Performance achieved by context models in the default condition, see Sec. 1. Diagrams are grouped by video streams of the HRI RoadTraffic dataset (stream I not shown due to space limitations). Dashed green curves: performance of signal-driven object detection alone. Yellow and blue curves: effects of including ground-truth-trained (yellow curves) and co-trained (blue curves) context models. A clear improvement can be observed for all streams in contrast to unaided signal-driven detection; co-trained context models achieve a performance very similar to to ground-truth-trained case.

5. Discussion

Our experiments show that, across all video streams, a markedly superior vehicle detection performance is achieved when coupling context models to signal-driven object detection. As the goal of this study is to show that co-training of context models is feasible in contrast to using ground-truth data, the performance in both cases must be compared. As can be clearly seen from the "default condition" experiment (Fig. 9), the replacement of ground-truth data by the co-training process has a very small influence on detection performance. To our mind, this decrease is more than compensated by the enormous reduction of required training data.

As it stands to reason that co-training can only work when the classifier generating the supervision signal data is of sufficient accuracy, we conducted the "impaired condition" experiment which used training data of reduced quality for signal-driven object detection. As can be expected, and as it is indeed seen in Fig. 10, this leads to reduced detection accuracy of signal-driven detection. Nevertheless, the coupling of co-trained context models results in a significant increase in detection performance, as seen from Fig. 10, although the performance seen in the default condition (Fig. 9) is not reached. The conclusion can only be that meaningful context models are still acquired in spite of the reduced quality of the supervision signal.

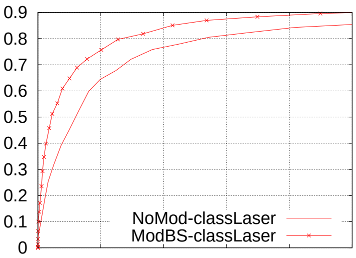

Figure 10: Performance achieved by context models in the impaired condition, see Sec. 1. Shown is the performance of signal-driven object detection alone (solid red curve), and in combination with co-trained context models (red curve with crosses). Apparently, feedback signals still cause a strong overall increase in detection performance, although it is inferior to that of the default condition.

Summary and conclusion. In this contribution, we benchmarked the performance of a hybrid vehicle detection system composed of a sliding-window-type object detector ("signal-driven detection") and a simple model of object-to-scene relations ("context models"), supplying a Bayesian prior distribution to the detector. Our focus was to reduce the overall amount o ground-truth data required for training the complete system; to this end, we compared the direct training of context models from ground-truth data to using the object hypotheses from signal-driven detection as training data, a procedure which we term "bootstrapping". Two bootstrapping scenarios were investigated, one where signal-driven detection was trained on ground-truth data, and another where signal-driven detection was trained on automatically generated vehicle data obtained by processing signals from a laser sensor. Our key findings were that, in all three cases, the obtained context models had a strong beneficial effect on detection performance, and that bootstrapping of context models was successful in all cases (although of slightly inferior performance). In particular, the case where context models were bootstrapped from signal-driven detection trained on automatically generated vehicle data deserves the highest attention, because it is essentially an object detection system that requires *no ground-truth data at all* to be trained to competitive performance.

To conclude, we have determined that, at least for vehicles, co-training of context models is a feasible option which will allow a huge reduction of cost and human effort when constructing powerful hybrid detection architectures. If a simple sensor exists that generates halfway reliable ground-truth data (at least in some defined situations), we may even speculate that the training of detection and context models could be done entirely without ground-truth data, by using the basic co-training techniques of this article.

References

[1] H Bischof, M Godec, C Leistner, B Rinner, and A Starzacher. Autonomous audio-supported learning of visual classifiers for traffic monitoring. 2009.

[2] A Blum and T Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*, 1998.

[3] C Desai, D Ramanan, and C Fowlkes. Discriminative models for multi-class object layout. In *International Conference on Computer Vision (ICCV)*, 2009.

[4] A Gepperth, J Edelbrunner, and T Bücher. Real-time detection of cars in video sequences. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV2005)*, pages 625–631, June 2005.

[5] A Gepperth, J Fritsch, and C Goerick. Cross-module learning as a first step towards a cognitive system concept. In *Proceedings of the First International Conference On Cognitive Systems*, 2008.

[6] A Gepperth, B Mersch, J Fritsch, and C Goerick. Color object recognition in real-world scenes. In JM de Sa, editor, *ICANN 2007, part II*, number 4669 in Lecture Notes in Computer Science. Springer Verlag Berlin Heidelberg New York, 2007.

[7] A Gepperth, S Rebhan, S Hasler, and J Fritsch. Biased competition in visual processing hierarchies: A learning approach using multiple cues. *Cognitive Computation*, 3(1):146–166, 2011.

[8] D Hoiem, A Efros, and M Hebert. Putting objects into perspective. *International Journal of Computer Vision*, 80(1), 2008.

[9] ibeo Automobile Sensor GmbH. Betriebsanleitung ibeo LUX Laserscanner, 2007.

[10] B Leibe, N Cornelis, K Cornelis, and L Van Gool. Dynamic 3D scene analysis from a moving vehicle. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007.

[11] C Leistner, A Saffari, PM Roth, and H Bischof. On robustness of on-line boosting - a competitive study. In *Online Learning in Computer Vision*, 2009.

[12] T Michalke, R Kastner, M Herbert, J Fritsch, and C Goerick. Adaptive multi-cue fusion for robust detection of unmarked inner-city streets. In *Proc. IEEE Intelligent Vehicles Symposium (IV'09)*, Xi'an, China, 2009.

[13] K Murphy, A Torralba, D Eaton, and WT Freeman. Object detection and localization using global and local features. In J Ponce, editor, *Toward Category-Level Object Recognition*, Lecture Notes in Computer Science. Springer, 2005.

[14] B Schiele P Dollar, C Wojek and P Perona. Pedestrian detection: A benchmark. In *CVPR*, June 2009.

[15] P Roth, H Bischof, D Skočaj, and A Leonardis. Object detection with bootstrapped learning. In Allan Hanbury and Horst Bischof, editors, *Proc. 10th Computer Vison Winterworkshop*, pages 33–42, 2005.

[16] P Roth, H Grabner, D Skocaj, H Bischof, and Aleš Leonardis. Conservative visual learning for object detection with minimal hand labeling effort. In *German Association for Pattern Recognition Yearly Symposium (DAGM)*, 2005.

[17] M Sonka, V Hlavac, and R Boyle. *Image Processing, Analysis and Machine Vision.* Chapman and Hall, 2 edition, 1995.

[18] M Szczot, I Dannenmann, and O Lohlein. Incorporating lane estimation as context source in pedestrian recognition task. In *ICPR*, pages 2628–2631, 2010.

[19] A Torralba. Contextual priming for object detection. *IJCV*, 53:2003, 2003.

[20] PA Viola and MJ Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518, 2001.

[21] J Vogel and K Murphy. A non-myopic approach to visual search. In *Computer and Robot Vision*, volume 0, pages 227–234, Los Alamitos, CA, USA, 2007. IEEE Computer Society.

[22] H Wersing, S Kirstein, B Schneiders, U Bauer-Wersing, and E Körner. Online learning for boostrapping of object recognition and localization in a biologically motivated architecture. In *Proc. Int. Conf. Computer Vision Systems ICVS. Santorini, Greece.*, pages 383–392, 2008.