Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Combining Multiple Classifiers and Context Information for Detecting Objects under Real-world Occlusion Patterns

Marvin Struwe, Stephan Hasler, Ute Bauer-Wersing

2013

Preprint:

This is an accepted article published in New Challenges in Neural Computation (NC2) Workshop. The final authenticated version is available online at: https://doi.org/[DOI not available]

Combining Multiple Classifiers and Context Information for Detecting Objects under Real-world Occlusion Patterns

Marvin Struwe[1], Stephan Hasler[2], and Ute Bauer-Wersing[1]

[1] University of Applied Sciences Frankfurt am Main, Germany
[2] Honda Research Institute Europe GmbH, Offenbach, Germany
{mstruwe,ubauer}@fb2.fh-frankfurt.de
stephan.hasler@honda-ri.de

Abstract. Current state-of-the-art detection approaches reveal a strong performance degradation with increasing object occlusion. Here we investigate different strategies to improve detection of occluded objects based on the analytic feature framework from [8] and compare the results in a car detection task. Motivated by an analysis of annotated traffic scenes we first describe a dedicated combination of classifiers to deal with the predominant car-car occlusion, and second, we propose a more general concept to handle vertical occlusion patterns. In a final test, depth information is used as additional local cue to reason about visible object parts. We report first improvements and discuss advantages and drawbacks of the individual approaches for further investigations.

Keywords: Object detection, Occlusion handling, Supervised learning

1 Introduction

Despite extensive research visual detection of objects in natural scenes is still not robustly solved. The reason for this is the large appearance variation in which objects or classes occur. A very challenging variation is occlusion which is caused by the constellation of objects in a scene. Occlusion reduces the number of visible features of an object but also causes accidental features. Existing object representations can deal moderately well with a low to medium level of occlusion and fail for stronger occlusions. The parts-based methods like [6, 7] aggregate local features in a voting manner and are usually trained with unoccluded views. During recognition they can handle arbitrary occlusion patterns, but require that sufficiently many features can still be detected. In contrast to this other methods like [1, 8] train a holistic object template in a discriminative manner. These methods focus resources on differences between classes. Because of this strong specialization on the training problem, these approaches show a stronger decrease of performance for occluded objects when trained on unoccluded views. However, in general the voting methods perform worse than the discriminative ones, whenever test and training set do not show such systematic

differences, as discussed in [13] and confirmed by the detection results in [2]. In this paper we investigate different strategies to improve the performance of a holistic discriminative method under stronger occlusion.

A common strategy to explicitly deal with occlusion is to make use of context information, i.e. to exploit knowledge about the possible constellation of objects. In [3, 12] Markov-Random-Fields are used to infer if neighboring features are consistent with a single detected instance of an object or have to be assigned to different ones. In this way both approaches can reason about relative depth of objects and produce a coarse segmentation. However, both methods require a time consuming iteration process over the whole input image. In [10] spatial relations are exploited more selectively by using the detections of larger and thus more easily detectable objects to narrow down search space for smaller, more difficult ones. We transfer this general concept to the occlusion case, where the difficult objects are the occluded ones.

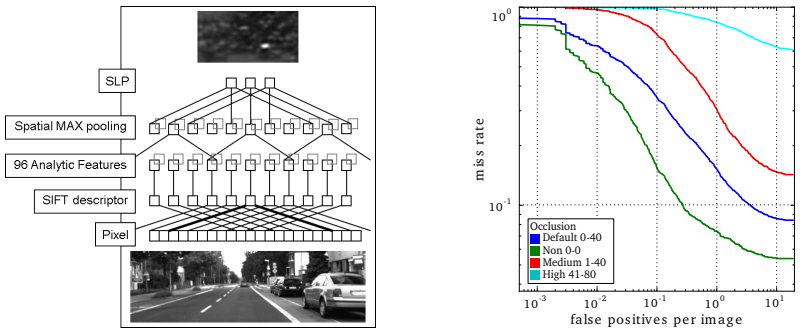
Besides instance-instance relations also knowledge about general occlusion patterns can be used. In [11] the authors explicitly take vertical occlusion at the image border into account. When objects are placed on a horizontal plane, vertical occlusion is a typical pattern also inside the image. Here we exploit this for car detection by using a dedicated classifier architecture.

Occlusion is related to the 3D relation of objects. A general cue of 3D information is depth. To check the physical plausibility of an object's position and size [4] or to segment and put attention to individual scene elements. So in [9] temporal differences between RGB-D(epth) views are used to discover movable parts for action representation. Here we integrate depth into the architecture for vertical occlusion to reason about visibility of features.

In Sec. 2 we shortly describe our basic holistic discriminative detection framework and show how it is influenced by occlusion in a car detection task. In the following sections different strategies to deal with occlusion are motivated and tested. First, in Sec. 3 we propose a conditional combination of classifiers for the predominant car-car occlusion setting. To deal with more general occlusion pattern we design a classifier architecture in Sec. 4 where the response of discriminative vertical parts detectors are integrated in a second stage. Finally, in Sec. 5 we make a first test how to exploit depth information in the 2-stage architecture, before drawing the conclusion in Sec. 6.

2 Analytic Feature Framework

Holistic discriminative approaches usually extract unspecific features and apply a powerful classifier directly on top. So the popular method proposed in [1] uses Histograms of Oriented Gradients (HOG) with a Support Vector Machine (SVM) and was shown to yield state-of-the-art performance in various detection tasks. In [8] we proposed the analytic feature framework (Fig. 1a) that puts effort in learning a more problem-specific feature representation and uses a simple classifier on top for discrimination. We could show that it provides competitive detection performance. For an input image first SIFT descriptors are computed



(a) (b)

Fig. 1: (a) Analytic feature hierarchy. SIFT descriptors are computed on a regular grid and matched to 96 analytic features. After a local maximum filter per feature the SLP templates are used in a convolutional step. Maxima in the final response map denote possible car locations. (b) ROC of C_{Std} for car detection scenario. The performance decreases strongly with the percentage of the cars' occlusion.

on a regular grid and then matched to a set of 96 analytic features. The analytic features are the result of a supervised selection process described in [5]. Next, a local maximum filter is performed per feature to enhance robustness against small translations. Finally the car template, which was trained with a Single Layer Perceptron (SLP), is shifted over the feature representation. The local maxima in the resulting response map denote possible car locations. To deal with cars at different distances we apply the framework on successively reduced image resolutions. For a pedestrian benchmark in [8] we could prove highly competitive performance of the analytic feature framework approach.

In [8] we used the framework for detection of front and back views of cars in real world traffic scenes. These image streams were taken under different weather conditions (sunny, rainy, overcast) and in different scene types (city, rural, industry, highway) and contained cars under all levels of occlusion. The final SLP car template was trained on unoccluded views only. We will refer to this reference system as C_{Std} throughout the paper.

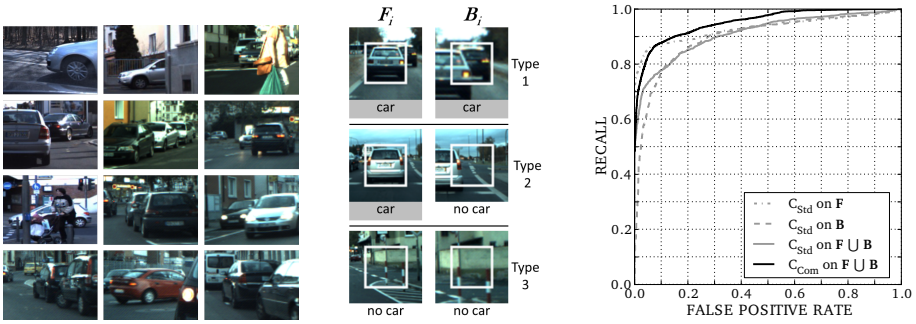
The results in Fig. 1b reveal a strong dependency of the performance of C_{Std} on the percentage of the cars' occlusion. For a false positive per image rate of 0.1 we get 70% of the cars with an occlusion between 0-40%. This pure detection performance is usually sufficient for a system that applies temporal integration (tracking). However for stronger occlusion the recall drops severely, which can no longer be compensated at system level.

3 Occlusion Handling Using Object-Object Relations

To get a better understanding of occlusion of cars in traffic scenes we counted the number of typical occluders and types of occlusion using ground truth data

Table 1: Counts of car occluders and occluded car parts for the ground truth data. In total 8796 out of 15514 cars are occluded, most of them by other cars.

Occluding object	#	Occluding object	#	Occluded part	#
Another car	7061	Pedestrian	70	Left	3730
Image border	2137	Traffic sign	31	Right	3124
Motor bike	82	Other/non-labeled	1125	Middle (only)	90



(a) (b) (c)

Fig. 2: (a) Typical occlusion examples. (b) Segment pair types. Each pair has a foreground segment $\boldsymbol{F}_i$ and a background segment $\boldsymbol{B}_i$. For $\boldsymbol{B}_i$ the first pair-type contains a labeled car occluded by another car, the second pair-type a random region next to a car, and the third pair-type a random region next to a false-positive of C_{Std}. For simplification we only use samples with occlusion at the left side and mirror examples with right occlusion to get more data. (c) C_{Std} shows a good performance for the foreground segments $\boldsymbol{F}$ while the result for the occluded cars $\boldsymbol{B}$ is significantly weaker. On the combined data set $\boldsymbol{F} \cup \boldsymbol{B}$, C_{Com} is in general much better than C_{Std}.

(see some occlusion examples in Fig. 2a). The result in Tab. 1 reveals that most cars are occluded by other cars. In [8] we exploited this fact and trained an additional classifier C_{Occ} on occluded cars and applied in the vicinity of cars already detected by C_{Std} only. This concept was inspired by [10] where the detection of small office objects was enhanced by predicting their spatial position relative to larger and more easily detectable objects.

We decided to use two separate classifiers instead of training C_{Std} also with occluded views, because we expected a decrease of performance of C_{Std} for unoccluded cars otherwise. The conditional application of C_{Occ} is necessary to avoid a strong increase in the number of false positives, which would be the result of the independent usage of both classifiers. We refer to the combined application of both classifiers as C_{Com}.

For a fast proof of concept, in [8] we tested this strategy on segmented car and non-car views. So we generated data pairs i, each having a **F**oreground segment $\boldsymbol{F}_i$ containing the occluder and the corresponding **B**ackground segment $\boldsymbol{B}_i$ containing something occluded. We refer to the set of all foreground/background segments with $\boldsymbol{F} = \{\boldsymbol{F}_i\}$ and $\boldsymbol{B} = \{\boldsymbol{B}_i\}$ respectively. The types of pairs that mimic all possible constellations in a scene are shown in Fig. 2b. We trained C_{Occ} on the background segments $\boldsymbol{B}$ including cars with an occlusion up to 80%.

Figure 2c confirms again that C_{Std} can cope substantially better with the familiar foreground segments $\boldsymbol{F}$ than with the occluded segments in $\boldsymbol{B}$. On the combined data set $\boldsymbol{F} \cup \boldsymbol{B}$ the classification has some intermediate quality but is clearly dominated by C_{Com}. For example, at a recall of 0.8 C_{Std} has a false positive rate of 0.13, while that of the combined curve is 0.04. This is a threefold reduction in the number of false positives.

The used segment dataset in [8] was simple in the way that the position and size of the occluded car was normalized, whereas in a real scene a strong variance can be expected relative to the position and size of the unoccluded car. Because of this it is not trivial to transfer the proposed concept to full scene detection. Furthermore the concept neglects general occluders. Thus we propose other new strategies to deal with occlusion in the following.

4 Split of the holistic car template

Tab. 1 reveals that most cars are either occluded on the right or left side. This vertical occlusion is caused by other cars, unlabeled walls, or the image border (see Fig. 2a), and causes a mismatch of the holistic car template used in C_{Std}. To improve the detection for this type of occlusion we used following strategy: We subdivided the holistic classifier into three vertical parts and trained each part-classifier with unoccluded car views. So each classifier is forced to make a more local decision about the presence of the car and is later not affected by occlusion of a different part. To integrate the responses of the part-classifiers we use their confidence values as input for an additional SLP which is trained with cars with occlusion rate 0-80%. The resulting two stage architecture is shown in Fig. 3b and will be referred to as C_{3Split}. The structure is equivalent to an MLP but instead of Backprop learning we use for each stage a different training set. In this way we enforce a stronger local decision and better generalization of each sub-segment.

The detection results in Fig. 4a show an improved performance of C_{3Split} compared to C_{Std} for occlusion rates of 1-40%, while there is a similar performance for occlusion rates of 41-80%. Unexpectedly, there is a strong gain for unoccluded cars. A possible reason might be that the vertical part templates of C_{3Split} are forced to make better use of their local information and thus find a more general car concept each. This hypothesis is underlined by the weights learned by the holistic SLP C_{Std}, that show a rather sparse contribution of a small set of analytic features at specific locations, while each part-classifier of C_{3Split} integrates all features at all positions in a much broader manner.

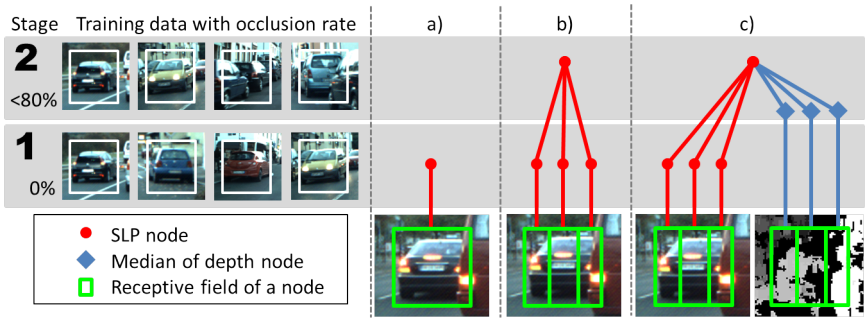


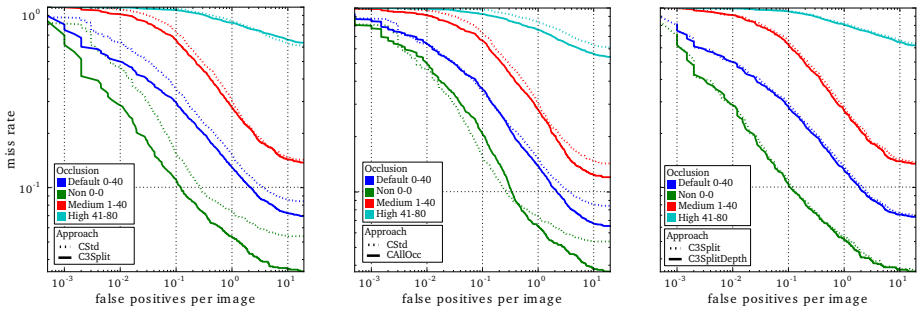
Fig. 3: Different detection architectures. (a) Architecture of C_{Std}. A holistic SLP template is trained on non-occluded car views. (b) Architecture of C_{3Split}. The three vertically subdivided SLP templates are trained on the same examples to (a). In the second stage, a separate SLP learns to combine the three confidence values of the first stage using non-occluded and occluded car views. (c) Architecture of $C_{\text{3SplitDepth}}$. The final SLP uses the median depth for each vertical car part as additional input.

Motivated by the result of C_{3Split} we splitted each vertical part further into two horizontal regions. However, C_{6Split} showed a much worse performance compared to C_{3Split} in a similar range as C_{Std}. In future more systematic analysis is necessary to investigate the effects of this two stage architecture.

C_{3Split} differs from C_{Std} in two ways: It uses an adapted classifier architecture and additional occluded training examples in the second stage. To verify that the improved performance of C_{3Split} is not simply caused by using different training data, we trained a holistic detector similar to C_{Std} but using the training data of the second stage of C_{3Split} with occlusion of 0-80%. We refer to this classifier as C_{AllOcc}. In Fig. 4b, C_{AllOcc} shows a similar gain as C_{3Split} for 1-40% occlusion. The performance for unoccluded cars is worse than C_{Std}. This indicates that the additional variation in the occluded training examples confuses the representation of unoccluded views, as we expected in Sec. 3. For 41-80% occlusion C_{AllOcc} outperforms C_{3Split} and C_{Std}, so the holistic approach can better make use of the remaining information in case of strong occlusion, maybe by directly representing the effect of the occlusion edge.

5 Additional use of depth information

C_{3Split} shows an improvement in detection performance for cars with an occlusion rate of 0-40%. For cars with stronger occlusion the performance is nearly the same as C_{Std}. The SLP in the second stage seems not to be able to deal with some occlusion cases by using only the confidence values of the sub-segments. A first analysis of the results shows the following problem: Very low confidence values in two of three sub-segments can result in a low confidence value at the classifier on top regardless of how good the confidence value on the third



(a) (b) (c)

Fig. 4: (a) Performance of C_{3Split}. C_{3Split} generally dominates C_{Std} at occlusion rates up to 40%, with an unexpected, strong gain for unoccluded views. (b) Performance of C_{AllOcc}. C_{AllOcc} shows a similar gain as C_{3Split} for 1-40% occlusion. For unoccluded views the performance is worse than C_{Std}. But C_{AllOcc} outperforms the other detectors on strongly occluded cars. (c) Performance of $C_{\text{3SplitDepth}}$. The way $C_{\text{3SplitDepth}}$ makes use of depth information does not lead to a significant improvement.

sub-segment is. There are two different scenarios where this constellation can happen. First, the object is a non-car object but a structure in one sub-segment accidentally generates a high confidence value, which can potentially happen quite often. Second, the object is a strongly occluded car and only visible in one of the three sub-segments. The approach is not able to distinguish both patterns and thus seems to require a strong response of two vertical parts at least. One possibility to overcome this limitation is to make use of additional context cues, e.g. stereo disparity. So the fact that the occluded part of a car will be closer to the camera than the visible part could be exploited.

We tested this hypothesis by retraining the SLP on the second stage of C_{3Split} using the median depth of the three vertical car parts as additional input. The resulting architecture is shown in Fig. 3c and referred to as $C_{\text{3SplitDepth}}$. The results in Fig. 4c show no significant improvement compared to C_{3Split}, so the $C_{\text{3SplitDepth}}$ seems not to be able to make use of the additional cue. An analysis of the top-level SLP weights shows that the median depth information is nearly unused. There are several possible reasons for this. Maybe the use of the median per part is a too strong reduction of information, or the use of the absolute median values does not produce a linearly separable pattern for different distances of the car and its occluder. We need a detailed analysis of the depth information to find a useful representation which can increase the detection performance.

6 Conclusion

In this paper we presented different alternatives to improve the detection of occluded objects. In a first test we combined two detectors, one for unoccluded

and one for occluded instances, in a conditional way, and reported enhanced performance in a car-car occlusion setting using segmented views. For the more general setting of vertical occlusion we proposed a vertical split of the holistic car template into three parts together with a second stage that learns typical combination of these parts responses for occluded car examples. This prototype outperformed the reference system for different levels of occlusion. However, there was only a very small gain for strongly occluded cars, but an unexpectedly large gain for unoccluded ones. We confirmed that the gain in performance is not simply caused by the additional use of occluded training examples and we also shortly tested an additional horizontal split of the vertical parts, which strongly reduced the performance. Finally, we integrated depth information into the vertical occlusion prototype to give the classifier an independent cue to reason about the response patterns of the vertical parts. However, the straightforward use of the median depth per part seems to be too coarse or inadequate to get a significant performance gain. In conclusion, the experiments already suggest some possible directions for further improvement. But much more in-depth analysis of the individual effects presented in this paper are necessary.

References

1. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. Proc. CVPR (2005) 886–893
2. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: A Benchmark. Proc. CVPR (2009) 304–311
3. Gao, T., Packer, B., Koller, D.: A Segmentation-aware Object Detection Model with Occlusion Handling. Proc. CVPR (2011) 1361–1368
4. Gould, S., Baumstarck, P., Quigley, M., Y. Ng, A., Koller, D.: Integrating Visual and Range Data for Robotic Object Detection. ECCV workshop M2SFA2 (2008)
5. Hasler, S., Wersing, H., Kirstein, S., Körner, E.: Large-scale Real-time Object Identification Based on Analytic Features. Proc. ICANN (2009) 663–672
6. Leibe, B., Schiele, B.: Interleaved Object Categorization and Segmentation. Proc. BMVC (2003) 759-768
7. Lowe, D. G.: Distinctive Image Features from Scale-invariant Keypoints. In: IJCV 60(2) (2004) 91–110
8. Struwe, M., Hasler, S., Bauer-Wersing, U.: Using Analytic Feature for the Destection of Occluded Objects. To appear in ICANN (2013)
9. Stückler, J., Behnke, S.: Hierarchical Object Discovery and Dense Modelling From Motion Cues in RGB-D Video. To appear in IJCAI (2013)
10. Torralba, A., Murphy, K. P., Freeman, W. T.: Contextual Models for Object Detection Using Boosted Random Fields. Proc. ICIP (2011) 653–656
11. Vedaldi, A.,Zisserman, A.: Structured Output Regression for Detection with Partial Truncation. Proc. NIPS (2009)
12. Winn, J., Shotton, J. D. J.: The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects. Proc. CVPR (2006) 37–44
13. Yi-Hsin, L., Tz-Huan, H., Tsai, A., Wen-Kai, L., Jui-Yang, T., Yung-Yu, C.: Pedestrian Detection in Images by Integrating Heterogeneous Detectors. IEEE Computer Symposium (ICS) (2010) 252–257