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

Using the Analytic Feature Framework for the Detection of Occluded Objects

Marvin Struwe, Stephan Hasler, Ute Bauer-Wersing

2013

Preprint:

This is an accepted article published in ICANN 2013. The final authenticated version is available online at: https://doi.org/[DOI not available]

Using the Analytic Feature Framework for the Detection of Occluded Objects

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. In this paper we apply the analytic feature framework, which was originally proposed for the large scale identification of segmented objects, for object detection in complex traffic scenes. We describe the necessary adaptations and show the competitiveness of the framework on different real-world data sets. Similar to the current state-of-the-art, the evaluation reveals a strong degradation of performance with increasing occlusion of the objects. We shortly discuss possible steps to tackle this problem and numerically analyze typical occlusion cases for a car detection task. Motivated by the fact that most cars are occluded by other cars, we present first promising results for a framework that uses separate classifiers for unoccluded and occluded cars and takes their mutual response characteristic into account. This training procedure can be applied to many other trainable detection approaches.

Keywords: Object detection, Supervised learning, Occlusion handling

1 Introduction

Despite extensive efforts the visual detection of objects in natural scenes is still not robustly solved. The current best approaches usually extract unspecific local features and apply a powerful classifier directly on top. So e.g. the combination of Histograms of Oriented Gradients (HOG) [2] with a Support Vector Machine (SVM) is reported to yield good performance in various detection benchmarks. In contrast to this are methods that put effort in learning a more problem-specific feature representation, on top of which a very simple classifier can be used for discrimination. An example for such a method is the analytic feature architecture proposed in [5], which showed high performance for large-scale identification of segmented object views. In this paper we show that such an architecture can also provide competitive results in detection tasks.

One main problem for systems acting in real world is that objects are often occluded. This affects currently used object representations in different ways. E.g. the methods that aggregate local features in a voting manner like [6, 7] are usually trained with unoccluded views. During recognition they can deal with arbitrary occlusion patterns as long as sufficiently many features can still be

detected. In contrast to this are the methods that train holistic object templates in a discriminative manner like [2, 5]. When training on unoccluded and testing on occluded views these approaches show a much stronger relative decrease in performance. The reason for this is the stronger specialization on the training problem by focusing resources on differences between classes. 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 [11, 3].

To exploit the benefits of discriminative approaches also for occluded objects, one could simply train them with occluded and unoccluded views. But this will likely reduce the performance for unoccluded views during testing. So more advanced processing is necessary.

One possibility is to exploit the relation between occluding and occluded object, which for natural scenes usually shows rather systematic patterns. The detection approach in [9] first searches for larger and thus more easily detectable objects and later exploits spatial relations to improve the detection of smaller, more difficult ones. This concept can be transferred to the occlusion problem. So one could train special detectors for different types of occlusion and exploit their mutual response characteristic in a scene.

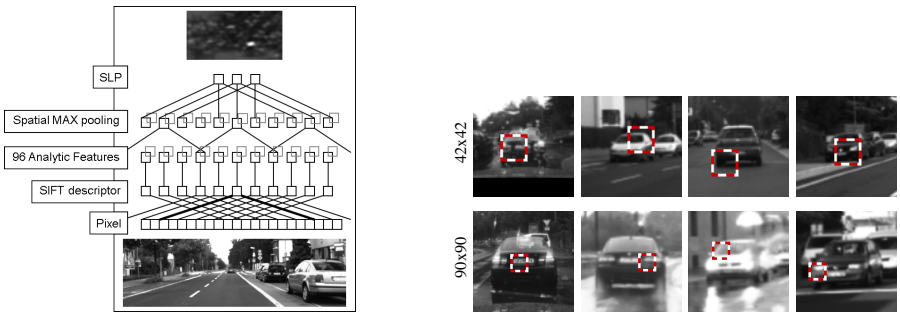
Other approaches that make use of object-object relations are presented in [10, 4], where Markov-Random-Fields are used to infer if neighboring features are consistent with a single detected instance or have to be assigned to different ones. In this way both approaches can reason about relative depth of objects and produce a coarse segmentation. However, in this paper we propose a more directed search for occluded objects, instead of using such demanding iterative processing over the full scene.

Also convolutional neural architectures were recently applied with great success on current recognition [1] and segmentation benchmarks [8]. The problem of occlusion, however, was not actively treated in these models so far. We propose a particular training procedure for occluded and unoccluded detectors that can be applied for these architectures similarly.

In Sec. 2 we outline how we adapted the analytic feature framework for detection tasks. After a short description of the traffic scene data used in our experiments, we evaluate the performance in Sec. 3. In Sec. 4 we propose a possible way to improve the detection of occluded cars and provide a first proof of concept on segmented car images, before drawing the conclusion in Sec. 5.

2 Adaptation of Analytic Feature Framework

We base our appearance-based detector on the real-time object identification framework in [5], which uses an attention mechanism to generate size normalized segments of the input object. Over the gray-scale segment first SIFT descriptors [7] are computed on a regular grid. Each descriptor is then matched to a set of 421 analytic features which are the result of the supervised selection process proposed in [5]. After this for each feature the global maximum is computed over the segment, in this way removing all spatial information. Finally a Single Layer



(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) Some analytic features for two template sizes.

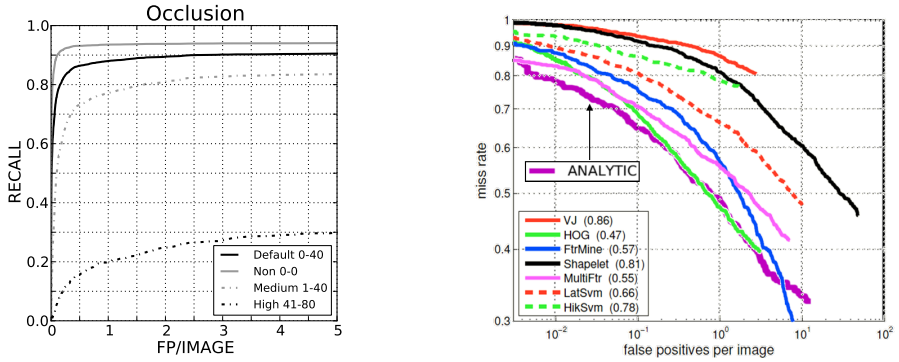
Perceptron (SLP) is used to separate the 126 objects in the 421-dimensional space. The approach is working robustly for full 3D rotation, even for untextured objects which are notoriously difficult for the standard SIFT approach [7].

The application of the existing framework for full scene object detection requires several adaptations. The rotation normalization of the SIFT descriptors is switched off because cars usually occur only upright. To speed up processing only 96 analytic features are used, where only features of the car class and not the background class were selected (see some examples in Fig. 1b). On the highest layer the local SLP template is shifted over the input image. This convolutional step is used to generate the car response map producing broad activation blobs for a car. The reason for this was the global maximum operation over features inside the template, which we had to replace with a local one to keep robustness against small translations. The resulting feature architecture is shown in Fig. 1a.

To deal with cars at different distances, we train analytic features and SLPs on three different segment sizes and use them on the largest image resolution, while the largest template is also used on successively reduced resolutions. This combined strategy improves detection performance because no compromise between minimal template size and most discriminative template size needs to be found, which is a common drawback of other detection approaches.

3 Detection Results

We decided to use the detection framework proposed in this paper to locate cars in real world traffic scenes. For this we equipped a car with a stereo camera and acquired different streams with a total length of 45 minutes covering different weather conditions (sunny, rainy, overcast) and scene types (city, rural, industry,



(a) (b)

Fig. 2: (a) ROC for our car detection scenario. The performance decreases strongly with the percentage of the cars' occlusion. (b) ROC for pedestrian benchmark (unoccluded, 50 pixels or taller). The analytic approach is on par with state-of-the-art approaches.

highway). For one frame per second we labeled typical traffic participants with a ROI and also roughly estimated their percentage of occlusion.

To evaluate the car detection performance we split the image streams into chunks of 30 seconds and used the odd chunks for the training of analytic features and SLP templates and the even chunks for testing. We decided to exclude car ROIs, whose width was more than twice their height (roughly 10% of the data), from training. In this way the use of smaller and square SLP templates was sufficient. So the templates were trained on segments of 42x42, 66x66, and 90x90 pixels (plus 18 pixels border at each side), into which the car was centered. Initially, the SLPs were trained to separate few thousand segments of unoccluded cars from a larger set of randomly chosen non-car segments. During 5 bootstrapping steps more negative examples were generated.

Please note that we used the disparity information available for our image data to reject implausible car candidates with simple hand-tuned rules on height-above-ground and physical-size. Finally, a local competition removed further weak hypotheses if they had a too strong overlap with more confident ones. For the input images of size 800x600 the GPU implementation of our framework runs with 10 frames per second on a mobile Geforce GTX580M.

The results for our car scenes are shown in Fig. 2a. We excluded cars with a height below 35 pixels and used the common 50% mutual overlap criterion between labels and detections. The curves reveal a strong dependency 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 some kind of temporal integration (tracking). However for higher percentages of occlusion the recall drops severely, which can no longer be compensated at system level. In the next

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

section we propose a special approach for detection of occluded cars.

To get a comparison with state-of-the-art methods we decided to apply our framework also to the pedestrian detection benchmark proposed in [3]. We evaluated the performance by doing 6-fold crossvalidation over the 6 streams and averaged the results. In contrast to this the competitors in Fig. 2b used all streams for testing and trained on other pedestrian data each. So the results are not 100% comparable. However, because the streams are quite different from each other and the overall label quality is not that high, there is no obvious advantage in using the streams for training. So Fig. 2b roughly shows that we are at least competitive to the popular HOG approach [2]. The main conceptual difference to HOG is that it applies an SVM directly on top of the local gradient histograms, while we use an additional projection to the analytic features and a simple SLP as classifier. Please note that because of the missing stereo information only the mutual overlap heuristic was used here.

4 Occlusion Handling Using Object-Object Relations

In this section we propose a method to increase the detection performance for occluded cars. Following our discussion in the introduction, for this we like to exploit object-object relations. Taking into account that most cars are occluded by other cars, as revealed by the analysis of the ground truth shown in Tab. 1, we decided to implement following simple strategy: In order not to decrease performance for unoccluded cars we will train a special classifier on occluded cars. This new classifier will be applied in the vicinity of already detected cars only. In a scene these initial car hypotheses are generated by the detection framework described before using the classifiers trained on unoccluded cars. The conditional application rule is necessary to avoid a strong increase of false positives (FP), which would be the result of the independent usage of both classifiers.

For a fast proof of concept, we decided to first test 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 described in Fig. 3.

$\boldsymbol{F}_i$	$\boldsymbol{B}_i$	
 car	 car	**1. Car occluding car:** The occluding car is inserted as positive example to $\boldsymbol{F}$ and the occluded car as positive to $\boldsymbol{B}$.
 car	 no car	**2. Car not occluding car:** The car is put as positive to $\boldsymbol{F}$ and a randomly chosen, car-free region in its vicinity as negative to $\boldsymbol{B}$.
 no car	 no car	**3. Car-free pairs:** In a real scene the initial detector will produce false positives. The FPs of our detection framework are inserted as negatives to $\boldsymbol{F}$ and a randomly chosen, car-free region next to each FP as negative example to $\boldsymbol{B}$.

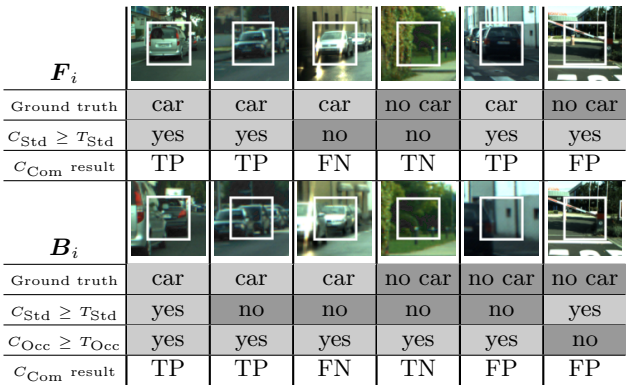
Fig. 3: Segment pair types. Each pair has a foreground segment $\boldsymbol{F}_i$ and a background segment $\boldsymbol{B}_i$. The positive examples (in gray) are generated from ground truth. For simplification we only use samples with occlusion at the left side and mirror examples with right occlusion to get more data. The classifier looks at the marked inner 42x42 pixel region of the segments into which the cars are fitted.

For the segment scenario we simply use the already trained SLP for a car size of 42x42 as initial classifier, which will be referred to as C_{Std}, and train the new classifier C_{Occ} on the background segments $\boldsymbol{B}$ of same size using cars with an occlusion up to 80%. The logic C_{Com} combines C_{Std} and C_{Occ} in a conditional manner and predicts the labels $L_{\boldsymbol{F}_i}$ and $L_{\boldsymbol{B}_i}$ for each pair using the code:

```
L_{F_i} = 'no car'
L_{B_i} = 'no car'
if C_Std(F_i) ≥ T_Std then
    L_{F_i} = 'car'
    if C_Std(B_i) ≥ T_Std then
        L_{B_i} = 'car'
    else if C_Occ(B_i) ≥ T_Occ then
        L_{B_i} = 'car'
```

So $\boldsymbol{B}_i$ is predicted as car, if either C_{Std} or the new classifier C_{Occ} reaches its corresponding threshold, **and** only if a car was found in the foreground segment $\boldsymbol{F}_i$ already. We have chosen a low FPR of 0.1 to avoid disturbing false alarms, which goes along with a recall of 0.84 ($C_{\text{Std}} = 0.37$, $C_{\text{Occ}} = 0.34$). Some classification examples are shown in Fig. 4.

To show the benefit of the combined logic, in Fig. 5a we compare the performance of C_{Com} with C_{Std} and in Fig. 5b with C_{Occ}. Please note that the combined approach depends on the two thresholds T_{Std} and T_{Occ}, and thus we have to evaluate the performance of C_{Com} for each combination of them. However, most of the resulting points in the ROC curve are dominated by a small set of other points. We reduced the amount of the points and only show in the plot a so called Pareto Front, which describes the set of optimal combinations.



$\boldsymbol{F}_i$						
Ground truth	car	car	car	no car	car	no car
$c_{\text{Std}} \geq \tau_{\text{Std}}$	yes	yes	no	no	yes	yes
c_{Com} result	TP	TP	FN	TN	TP	FP
$\boldsymbol{B}_i$						
Ground truth	car	car	car	no car	no car	no car
$c_{\text{Std}} \geq \tau_{\text{Std}}$	yes	no	no	no	no	yes
$c_{\text{Occ}} \geq \tau_{\text{Occ}}$	yes	yes	yes	yes	yes	no
c_{Com} result	TP	TP	FN	TN	FP	FP

TP - true positive
TN - true negative
FP - false positive
FN - false negative

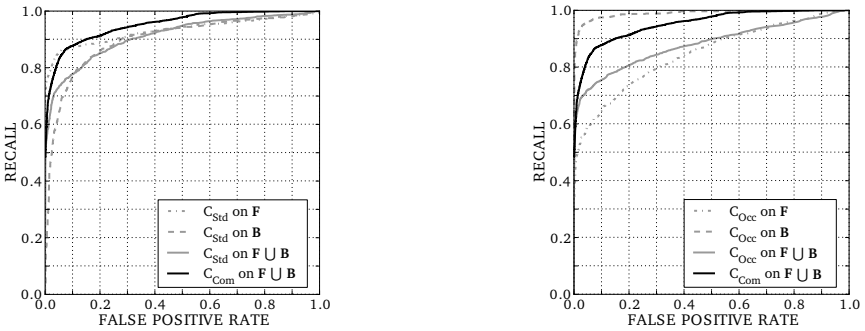
Fig. 4: Pair classification examples. For each foreground sample $\boldsymbol{F}_i$ we show the ground truth label, the decision of C_{Std}, and the result of the combined approach C_{Com}. For $\boldsymbol{B}_i$ we additionally show the decision of C_{Occ} because C_{Com} depends on both classifiers and on the result for $\boldsymbol{F}_i$. Dark gray is used for 'no car' labels and responses below threshold, light gray for the opposite. The conditional logic can correct FPs of C_{Occ} (4^{th} column), but in rare cases also prevents correct detections (3^{rd} column). The classifiers look at the marked inner 42x42 region.

Figure 5a 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.

In Figure 5b, C_{Occ} shows a very good performance on the occluded segments $\boldsymbol{B}$, for which it was trained. However, the performance for the unoccluded cars in $\boldsymbol{F}$ is much worse. One reason for this might be that C_{Occ} specialized too strongly on the edge that is caused by the occluder and which is not present in the unoccluded examples. Also in comparison to C_{Occ}, C_{Com} shows a substantially improved performance on the full data ensemble.

5 Conclusion

In this paper we presented a new object detection framework, which is based on the analytic feature representation originally proposed for object identification. We have shown the detection performance of the approach on a public pedestrian detection benchmark and evaluated on our own benchmark how strong occlusion effects the detection of cars. Motivated by these results and by an analysis of typical occlusion causes we proposed a new combination of detectors that takes the occlusion of cars by other cars into account. In a pre-study we successfully showed the benefit of this approach on segmented car views. The next step



(a) (b)

Fig. 5: Comparison of C_{Com} with C_{Std} and C_{Occ}. (a) 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}. (b) C_{Occ} performs very good on $\boldsymbol{B}$ but has strong problems on the unfamiliar occluders $\boldsymbol{F}$. C_{Com} also clearly outperforms C_{Occ} on $\boldsymbol{F} \cup \boldsymbol{B}$.

is to exploit the same principle also in full-scene detection and to other object relations, like persons occluded by other objects, or more general occlusion cases.

References

1. Ciresan, D. C., Meier, U., Gambardella, L. M., Schmidhuber, J.: Deep, Big, Simple Neural Nets for Handwritten Digit Recognition. In: Neural Computation 22(12) (2010) 3207–3220
2. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. CVPR (2005) 886–893
3. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: A Benchmark. CVPR (2009) 304–311
4. Gao, T., Packer, B., Koller, D.: A Segmentation-aware Object Detection Model with Occlusion Handling. CVPR (2011) 1361–1368
5. Hasler, S., Wersing, H., Kirstein, S., Körner, E.: Large-scale Real-time Object Identification Based on Analytic Features. ICANN (2009) 663–672
6. Leibe, B., Schiele, B.: Interleaved Object Categorization and Segmentation. BMVC (2003) 759-768
7. Lowe, D. G.: Distinctive Image Features from Scale-invariant Keypoints. In: IJCV 60(2) (2004) 91–110
8. Schulz, H., Behnke, S.: Learning Object-Class Segmentation with Convolutional Neural Networks. ESANN (2012) 151–156
9. Torralba, A., Murphy, K. P., Freeman, W. T.: Contextual Models for Object Detection Using Boosted Random Fields. ICIP (2011) 653–656
10. Winn, J., Shotton, J. D. J.: The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects. CVPR (2006) 37–44
11. 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. ICS (2010) 252–257