Chapter 8

Discussion and outlook

The work presented in this thesis does not describe a single object detection system but improvements to several aspects and subproblems of object detection. The goals have been to simplify the process wherever possible by neural learning methods, and to reduce seemingly different aspects of object detection to common principles.

8.1 Applications

Some of the results described in this thesis can be used directly to improve existing object detection systems. Due to the focus of my research group, these systems are concerned with object detection in traffic scenarios. Depending on the type of traffic scenes that are considered, some assumptions about the image content can be made. For example, well-marked lane borders can always be expected in highway traffic scenes, road models can be expected to hold, and the types of behaviorally relevant objects that can conceivably appear in an image are limited. In contrast, inner-city scenes are much more complex and unpredictable, road models hold only locally, if at all, and the recognition of a variety of object types is necessary for assessing a situation correctly.

Highway traffic scenes are therefore a strongly simplified scenario, and any object detection system should work reliably in this case. However, the object detection systems developed at my research group are intended to perform well even in more difficult surroundings.

Typical objects to be detected are cars, traffic signs and lane borders. The efforts described below are already implemented or currently under development. They are briefly described in order to demonstrate that the results obtained in this thesis have direct consequences for technological applications.

Of equal importance are the possibilities for further academic research that are suggested by this thesis; they are sketched further below.
8.1.1 Trainable initial detection

Using the results from chapter 6, it is possible to construct fast initial object detectors that can be trained by examples. For this purpose, one simply uses the sparse convolutional neural networks (SCNNs) from chapter 6, keeping the network complexity (i.e., the number of independent parameters) as low as possible, thus ensuring the fastest possible execution speed. The limited network capacity may lead to misses or false detections; nevertheless, if the threshold that defines the decision of the neural network (NN) is kept low, the detection can be biased such that all incorrect classifications are false detections. This is not a problem since one can use SCNNs of increased capacity as object classifiers that classify only those ROIs that have been found by the initial detection. In this way, the increased computational effort of simulating more complex networks is only expended if it is really necessary, thus keeping the total processing time low.

It is imperative not to train the object classifiers on the same data as the initial detection SCNN. Instead, new training datasets must be constructed containing all of the previous positive examples and the false detections produced by the initial detection SCNN as negative examples.

This framework can be extended to include the simultaneous initial detection of several object classes using the feature base learning technique from chapter 6. Although each object class requires a separate object classifier, this does not increase the computational load greatly because object classifiers are only applied to ROIs provided by the initial detection. Since initial detection is performed using a common feature base, processing time is approximately independent of the number of object classes provided that feature base learning itself gives satisfactory results.

The facilitation of initial detection design is a general result of this thesis: in the following section, a specialized application is presented which directly uses the results described in chapters 4, 6 and 7.

8.1.2 Initial detection of cars and traffic signs

In order to make detection maximally robust, a project is currently pursued to integrate histogram-based initial detection and initial detection with SCNNs. For the purpose of performing initial detection using the SOE features described in chapter 4, the whole image is partitioned into receptive fields (RF) in a way that is similar to the partitioning of ROIs in chapter 4. By using NN classifiers of various fixed sizes that are optimized for speed by magnitude-based pruning,1 the whole image can be scanned for objects very quickly. The NN classifiers are identical to those described in chapter 7.

The SOE features are additive: this means that one can compute SOE features at coarser spatial scales directly from the SOE features of finer scales.

---

1It was shown in chapter 7 that pruning is an acceptable optimization method for the car classification problem. Since traffic sign classification gives comparable classification errors, it is assumed that pruning can be applied there, too.
Figure 8.1: Initial detection (for clarity, only car detection is shown) by a combination of simple SCNN classifiers (upper right) discussed in chapter 6 and NN classifiers that rely on the SOE features (upper left) described in chapter 4. For the latter, the image is subdivided into small receptive fields (indicated by the grid structure) in each of which SOE features are computed. The NN classifier is applied at various scales and all positions within the image that are multiples of the receptive field size. In contrast, the SCNN classifier can be replicated over the whole image (also at several spatial scales) using finer intervals that are defined by the structure of the SCNN. The application at all possible locations is indicated by arrows in the upper row of images. The "AND"-operation symbolizes that ROIs must be detected by both methods in order to be considered.

Therefore, no new information can be gained by computing SOE features at multiple scales, and the use of differently sized classifiers for multiscale object detection is thus justified.

The results of initial detection using the SOE features are compared to those obtained from the trainable initial detection using SCNNs. Only those objects that are found by both methods are considered to be detections, although a more flexible combination strategy may be favorable. Detections are then examined by object classifiers implemented by a more complex SCNN in case of cars, and by a correlation-based classifier in the case of traffic signs. Please see fig. 8.1 for an overview of the initial detection architecture. Initial results are very promising both from the point of view of detection accuracy and real-time capability.

8.1.3 Classification of lane borders

The correct detection of lane and road borders can simplify the analysis of traffic scenes considerably. On the one hand, certain image regions can be excluded
from further analysis, and on the other hand, assumptions can be made about the object classes to be expected on and off a road.

Figure 8.2: Lane detection by combining a specialized initial detection algorithm and an SCNN classifier. The initial detection produces points (indicated by small black boxes) which are conjectured to lie on a lane border. The classifier considers a small ROI around those points for its decision. Confirmed lane borders are indicated by bright boxes.

Since lane and road borders are essentially white lines, the initial detection task is comparatively simple; in contrast, real-time constraints tend to be much more restrictive because other, more time-consuming detection tasks must also be performed in real-time. Computationally demanding feature extraction is thus out of the question as well as classification with complex NNs. For this reason, SCNNs are employed for classifying ROIs; the ROIs are found by an initial detection system that was designed by an industrial partner. The advantage in this case lies in the fact that no feature extraction needs to be performed, and that the classification is fast enough to process more than 100 ROIs per image under real-time conditions (approximately 25 images per second). A typical image with correctly classified ROIs is shown in fig. 8.2.

8.2 Opportunities for further research

Beyond the mere applications that can be realized and that were described in the previous section, a number of interesting research topics suggest themselves based on results presented in this thesis. Without going into details too much, some of the possibilities are outlined.

8.2.1 Extensions of the SCNN model

The SCNN model is considerably simplified compared to the original CNN proposal [83]. This is quite intentional; however, experiments need to be conducted to find out if extensions to the SCNN model can improve its capabilities. In this
respect, the inclusion of downsampling layers (as in [83], also sketched in fig. 3.2) is worth investigating, and furthermore the possibility of shortcut connections that bypass one or several layers.

8.2.2 Research of new unsupervised learning terms for the SCNN model

The SCNN learning rule presented in chapter 6 uses an approximation to the nonlinear subspace PCA learning rule. Two issues are of interest in this respect: First of all, it would be interesting to know if other learning rules could be used. For example, it is conceivable that rules associated with independent component analysis (ICA, see, e.g., [57]) may yield superior results. On the other hand, the question is of interest how features can be generated that are correlated with the object class. Stated in another way, it should be found out if certain local features are characteristic of a certain object class per se, and, if so, what learning rule is needed to obtain them. Some interesting results have been published recently in this direction [49, 109] which could serve as a starting point for investigations.

8.2.3 Comparison of designed and learned feature extraction schemes

The SCNN model claims to compute meaningful and diverse object features as a part of its learning algorithm. It would be interesting to find out how the SCNN model (possibly extended using ideas from the previous section) compares to classifiers using feature extraction schemes tailored to suit their classification problems. This comparison has been done in chapter 6 for the car classification task with good results. Nevertheless, an investigation using a larger number of classification problems as benchmarks is desirable.

8.2.4 Ensemble learning with members of Pareto fronts

Multi-objective structure optimization of object classifiers as presented in chapter 7 produces no single best NN solutions but Pareto fronts of solutions. Each member of a Pareto front is optimal w.r.t. a certain trade-off between all optimization objectives. When using the objectives of speed (number of NN connections) and classification accuracy, this means that members of the Pareto front will vary from very small NNs with suboptimal classification accuracy to large NNs with optimal accuracy. Recently, ensemble learning methods have become popular [111]; it could be investigated if classification results can be improved using the AdaBoost method [111] with selected NNs from a Pareto front as an ensemble. Depending on the ensemble size, this would slow down classification; in addition, large NNs may be ensemble members, leading to a further increase in computational complexity. This need not necessarily be a problem if initial detection performs sufficiently well, which means that the number of objects to be examined by an ensemble classifier stays within reasonable limits.
AdaBoost has been shown to be beneficial w.r.t. classification accuracy and generalization ability; an empirical observation that has not yet been sufficiently explained is the fact that AdaBoost tends to be very resistant to overfitting.

### 8.2.5 Saliency-based object detection

The saliency map model presented in chapter 5 is currently not suitable for real-time applications but very interesting for addressing general issues of visual learning.

![Figure 8.3](image.png)

Figure 8.3: Example of a difficult detection task: The head of the pedestrian cannot be discerned from the background. This creates problems for detection strategies that attempt to find whole objects, in this case pedestrians.

Of particular interest to me is the issue of detecting object parts; this problem arises whenever the initial detection of whole objects is difficult. A good example is the problem of pedestrian detection (see fig. 8.3). In such situations, objects could be detected by finding parts of them first, and then inferring the position of whole objects from the positions of their parts.

Trainable saliency maps may prove to be an important tool for addressing this and related problems. The following interdependent questions seem—in my opinion—to be of special relevance:

- **Can object part categories be learned from salient object regions?**
  
  This can be tested by calculating the saliency map not for a whole image but only for ROIs containing training examples. By the procedure described in the discussion of chapter 5, a sequence of salient regions can be obtained in decreasing order of saliency. Then, after estimating the scale of the salient region (this should be possible by determining the scale of the strongest local feature map response), an unsupervised learning algorithm can be applied to the salient region that attempts to cluster the image content into categories. Self-organizing maps [78, 123] or similar learning methods may be of relevance at this point.
• *Can an object be robustly classified by classifying the learned parts?* Provided it is possible to learn object part categories, the learned categories could be used to characterize an object; of special interest is the question whether this is also possible when the object is partly occluded.

• *Can objects be found by detecting previously learned object parts?* Again, provided that object parts have been attributed to salient areas within objects, it is a question whether those areas are salient enough (or can be made so by learning) to be detected quickly in a cluttered image. A powerful learning algorithm for saliency maps will be crucial to these investigations.