Traffic sign detection and recognition with convolutional neural networks

ALEXANDER HANEL[1] & UWE STILLA[1]

Abstract: Traffic sign detection and recognition in street scene images of a vehicle camera allow to localize the traffic signs in a street scene image and to classify them with regard to their semantic meaning for the car driver. In this contribution, a method is described to detect traffic signs in a street scene image by evaluating image patches, sampled by a sliding window approach, with a convolutional neural network detector. Robust shape fitting is performed on the image patch of a positive detection to obtain the exact position of the traffic sign shape in the patch. A second convolutional neural network is applied to the image patches centred on the fitted shapes to classify the meaning of these traffic signs. The networks are trained and tested with samples of traffic signs and other street scene objects from the GTSRB and GTSDB datasets. The results have shown that models for traffic sign detection and recognition can be trained with an overall accuracy of more than 90 % obtained for a test set. The position of a traffic sign, known from shape fitting, has been shown to be an important a-priori knowledge to select the appropriate image patch to ensure a high accuracy of the subsequent traffic sign recognition.

1 Safer roads by traffic sign warnings

Modern advanced driver assistance systems are an important contribution to increase the safety on roads. An assistance system using traffic sign recognition can warn a car driver against a speed limit by means of optical hints (EURO NCAP 2013), for example. As well as a human driver, an assistance system needs to extract information from the environment around the car. Cars can therefore be equipped with sensors to perceive the environment. As the upcoming drive way and the traffic signs applicable for that way are in front of the vehicle, a front-looking sensor is suitable for traffic sign detection and recognition (e.g. LARSSON & FELSBERG 2011). Due to small costs compared to a Radar or LiDAR sensor (ETKBMW.COM), often an optical mono-camera is the first choice for this task.

Such a camera observing the environment acquires typically an image of the street scene including the street, traffic signs, buildings, traffic participants and further static and dynamic objects (Fig. 1). There exist algorithms for object detection which use the complete image of the street scene as input (e.g. REDMON et al. 2016). Other algorithms require an image patch centered on the object which should be detected (red rectangle in Fig. 1) as input (e.g. SERMANET & LECUN 2011). In the second case, the input image patch can be obtained by a sliding window approach sampling from the street scene image, for example.

Traffic sign detection and traffic sign recognition have different objectives (ZHU et al. 2016): the objective of traffic sign detection is to decide whether the input image, or the image patch, shows

[1] Technical University of Munich, Photogrammetry & Remote Sensing, Arcisstraße 21, D-80333 Munich, email: [alexander.hanel, stilla]@tum.de

a traffic sign or another arbitrary object. In contrast, the objective of traffic sign recognition is typically to classify the meaning of a traffic sign, given an image patch from which it is known to show a traffic sign (e.g. German Traffic Sign Recognition Benchmark dataset, STALLKAMP et al. 2012).

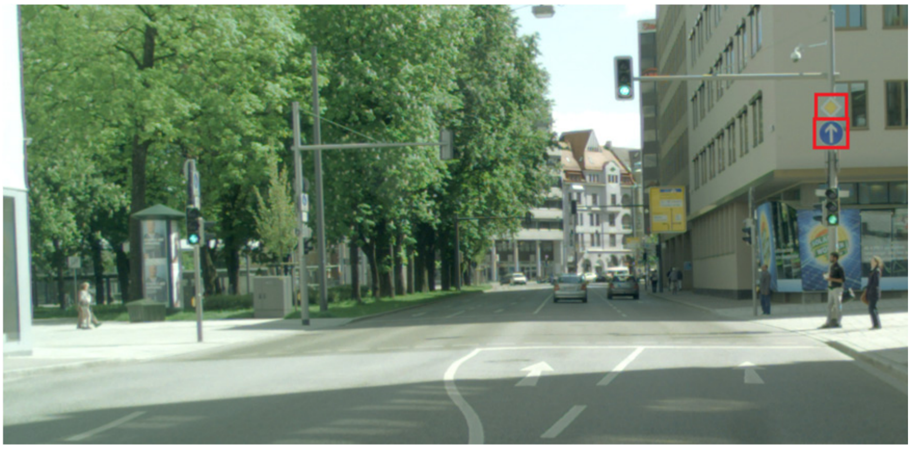

Fig. 1: Example image of a street scene recorded by a front-looking mono-camera. The image is an excerpt from the Cityscapes dataset (CORDTS et al. 2016), which can be used to train machine learning algorithms in the domain of advanced driver assistance systems, for example for traffic sign recognition

In a recent work, ZHU et al. (2016) propose a method for simultaneous detection and recognition of traffic signs using a neural network-based deep learning approach. Their method provides the bounding box, a pixel-wise mask and the class label for each detection. Several examples (in supplementary material provided by these authors) show that the bounding boxes of partly occluded traffic signs tend to be smaller than the sign in the image is. Using the pixels of the mask to extract the shape contour of the traffic sign will therefore be prone to errors.

HANEL & STILLA (2018) have recently proposed a method to calibrate a vehicle camera on public roads, where the exact shape of a traffic sign in the image has to be known precisely. Their method derives image points along the shape of a sign and the corresponding metric 3D position, for example known from governmental regulations, as reference points for calibration.

In this contribution, a method for traffic sign detection and recognition is proposed, which delivers the correct shape of the sign in the image simultaneously. A sliding window approach is used to sample image patches from a street scene image. The samples are evaluated by a traffic sign detector, which is realized by a neural network-based deep learning approach. Deep approaches have shown to provide a higher accuracy than shallow learning approaches, as for example ZHU et al. 2016 mention. Mean shift clustering is used to reduce multiple detections of the same traffic. Such multiple detections with slightly different positions and sizes are typical for a sliding window approach (COMASCHI et al. 2013). The shape of the traffic sign in the image is obtained by fitting an appropriate geometric primitive (e.g. an ellipse for circular traffic signs) with RANSAC followed by a least-squares adjustment. The extracted shape is used to refine the image position of the patch, which is then used by another deep learning approach to classify the meaning of the traffic sign.

2 Stepwise traffic sign detection and recognition

By the sliding window approach, image patches are taken from an image with a higher geometric resolution (called ‘complete image’ in the following). In the scope of this contribution, the complete image is typically a street scene image. Each patch shows a different part of the complete image; the patch is shifted over the complete image and the complete image is resized to consider traffic signs at different image positions and with different sizes in the image. The deep network architecture (Fig. 2) of the traffic sign detector uses such an image patch as input sample. In contrast to architectures, which use the complete image as input, the sliding window approach allows to reduce the computational effort when tracking traffic signs in an image series by evaluating only the local neighbourhood around the predicted image position of a traffic sign in a subsequent image in contrast to evaluating the complete subsequent image.

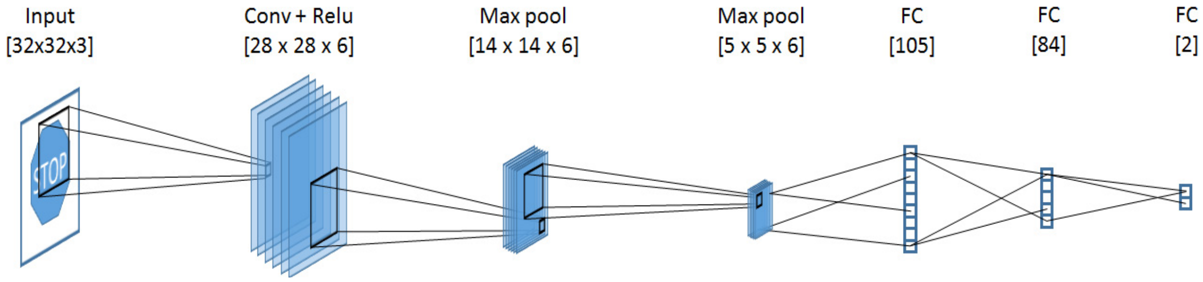

Fig. 2: Architecture of the convolutional neural network used to distinguish between *traffic signs* and *other objects* shown in an image patch taken from a street scene image.

The traffic sign detector classifies each sample in either the class *traffic sign* or *other objects*. The architecture of our network as described in the following is a variation of the traffic sign detector architecture described by WU et al. (2013): the input sample with a fixed resolution of 32 x 32 pixels and RGB color channels is first convolved with different kernels to extract features. The “ReLu” activation function ensures the non-linearity property of the network to prevent that the network behaves just like a single-layer perceptron. Max pooling in the second and third layer reduces the resolution of the feature maps created by the previous convolutions. Dropout is applied to the third layer to avoid overfitting. The following layers are designed to classify the input patch based on the extracted features: the output of the feature extraction layers is flattened, followed by three fully-connected (FC) layers in which the number of neurons is reduced step-by-step. The last FC layer has two output neurons, one for each of the two classes to distinguish. Softmax loss is used. In contrast to by WU et al. (2013), we use no branching (for details on branching see the description of the traffic sign recognition architecture), but increase the depth of the classification part. ZHU et al. (2016) have mentioned that deeper networks perform better. According to these authors, branching could also increase the performance, but at the cost of a higher computational effort.

The detector model is trained with a set of positive and negative samples. Image patches showing one traffic sign are called positive samples, while image patches showing arbitrary objects like buildings or vegetation are called negative samples. The class of an image patch with unknown content is predicted after training by feeding the patch together with the trained model into the neural network. Output of the prediction are the probabilities for both classes; the image patch is assigned to the class with the higher probability.

Image patches from slightly different positions and sizes in the complete image are likely to show the same object. Therefore, it can be expected that several of such neighboring patches will be detected as *traffic sign*. To link all detections belonging to the same traffic sign, the detections are grouped to a single detection by mean shift clustering (FUKUNAGA & HOSTETLER 1975). Each detection is hereby represented by the central point of the image patch. Mean shift clustering groups the central points of all detections into several groups using the Euclidean distance as criterion to distinguish different groups. The size of the image patch belonging to a single detection is determined by the mean size of the patches of all contributing detections. An advantageous property of mean shift clustering is that it does not require the number of clusters to be determined manually, being therefore able to handle a varying number of traffic signs in different complete images.

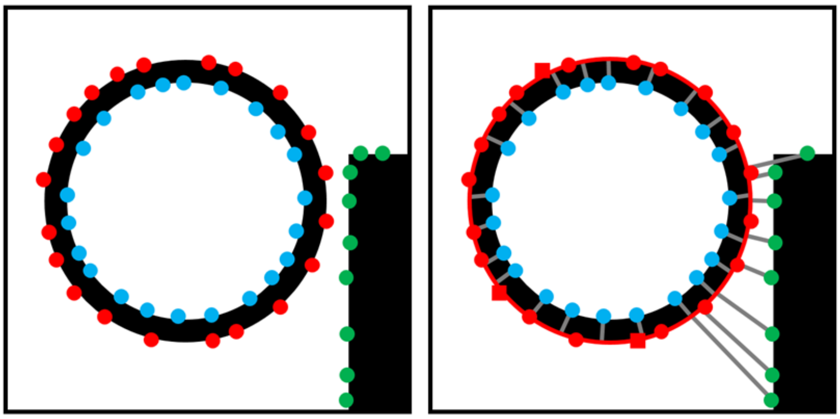

Fig. 3: Left: Edge of a traffic sign (black circle) and a background object (black rectangle) in a binary image patch. Contour points (red, blue, green balls) are along the edges of these objects. Right: Ellipse (red circle) fitted with sample points (red squares) selected from the contour points. The distance (grey lines) threshold for each contour point from the ellipse defines the inlier set used for the final-least-squares ellipse fit

As next step, the shape of the traffic sign in the image patch of each single detection is extracted by ellipse fitting. Result of the shape extraction is the position, size and orientation of the ellipse in the complete image used as precise location of a traffic sign prior to classifying the meaning of the sign. Further, mean shift clustering and shape fitting could be used to recognize false positive detections (e.g. by a small consensus set, high eccentricity of the ellipse, ...). The method for shape extraction is in the following described for circular traffic signs to allow analyses on the potential of this approach before further research is done. The method assumes that an image patch shows a traffic sign completely. To ensure this, the part of the complete image the patch covers can be increased. The ellipse is selected as geometric primitive as a real-world circle is projected to an ellipse in an image in general (e.g. AHN et al. 1999). The ellipse is fitted to groups of contour points extracted from the binary image patch. The image patch is binarized by applying an absolute global intensity threshold to it to separate objects with different unique intensities. The algorithm of SUZUKI & ABE (1985) is used to extract a group of contour points

(Fig. 3 left) along the edges of each image region with unique intensity in the binary image. Groups with a small number of points are neglected, assuming that the traffic sign is the dominant region with a high number of points compared to other groups in the patch. The RANSAC algorithm (FISCHLER & BOLLES 1981) is applied to randomly chosen contour points to select iteratively the consensus set of contour points belonging to the edge of the traffic sign shape. The inliers are defined by a maximal distance of the contour points to the ellipse in each iteration (Fig. 3 right). The largest set of inliers in all iterations is used as consensus set. The number of iterations is chosen to have chosen at least one set of sample points without outliers with a probability of 99 %. The final ellipse parameters are estimated in a least-squares-adjustment using the consensus set.

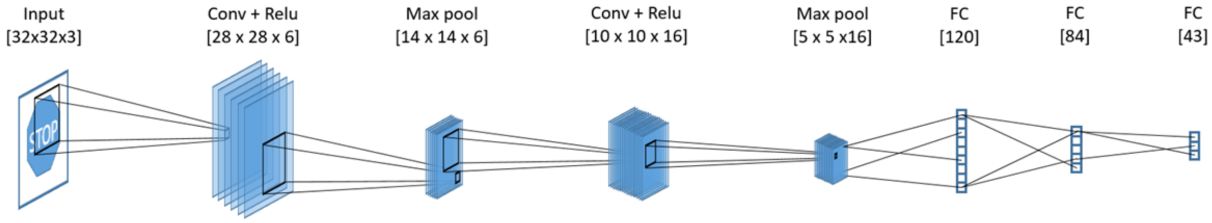

Fig. 4: Architecture of the neural network used to recognize the semantic meaning of a traffic sign shown in an image patch. In total, 43 different meanings are considered.

To recognize the semantic meaning of a traffic sign shown in an image patch, another machine learning method based on a convolutional neural network is used (Fig. 4). Our network is adapted from SERMANET & LECUN (2011). Its architecture can be divided in a feature extraction and a classification part as well. In the feature extraction part of our network, the so-called *skip architecture* used by these authors is left out. Thereby, the output of a layer would be branched and fed to a later layer by skipping some layers in between, while another branch uses all layers in between. More global features of traffic signs could be extracted by using branching. Instead, we increase the depth in the classification part, i.e. we increase the number of FC layers to three. The number of output neurons in the last layer is determined by the number of classes of traffic sign meanings in *The German Traffic Sign Recognition Benchmark* dataset (STALLKAMP et al. 2012).

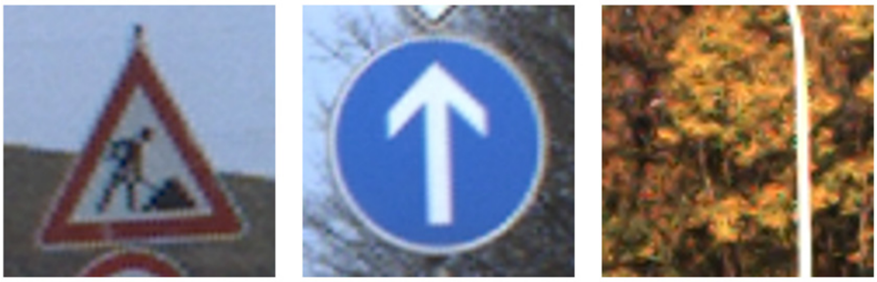

Fig. 5: Positive samples (taken from the GRSRB traffic sign recognition dataset). Left: Triangle-shaped construction warning sign in the shadow. Center: Circle-shaped go straight sign in sunny light. Right: Negative sample showing vegetation (extracted from street images from the GTSDB dataset)

3 Dataset and experiments

The models for the detector are trained with a set of sample patches showing one traffic sign or other objects, respectively. Separate models are trained for the following groups of traffic signs with different shapes and dominant colours: prohibitive signs (circle, red), mandatory signs (circle, blue), danger (triangle, red) and others (e.g. right of way, give way). One joint model for traffic sign recognition is trained with samples showing traffic signs with various meanings. The traffic sign samples (example in Fig. 5 left and centre) are taken from *The German Traffic Sign Recognition Benchmark* dataset (STALLKAMP et al. 2012). This dataset provides traffic sign image patches in various daylight illumination situations (e.g. sunny, shadowy) with a roughly frontal view on them, which would be also typical for a front-looking vehicle camera. The negative samples (example in Fig. 5 right) are extracted by randomly sampling image patches from street scene images from *The German Traffic Sign Detection Benchmark* (GTSDB) dataset (HOUBEN et al. 2013). The order of the samples is randomly shuffled, a subset of 80% of the samples is used for training and validation. The remaining subset of 20% is used for tests. The training is performed with a learning rate of 0.001 and a dropout rate of 0.5. 80 epochs are used in training, until there is no remarkable decrease in the loss anymore. The hyperparameter values have been obtained by hand-tuning. The proposed method is tested with 900 street scene images of the GTSDB dataset, for which the ground truth position and meaning of traffic signs are known.

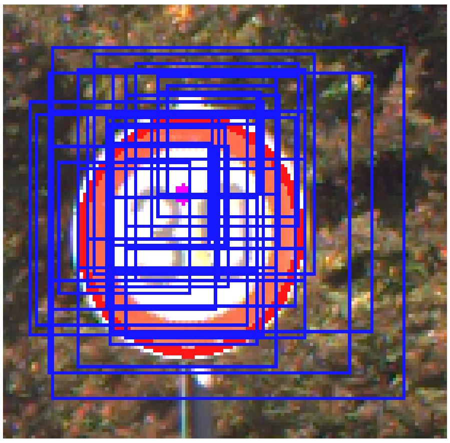

Fig. 6: Multiple detections (blue rectangles) for one traffic sign. The central point (pink dot) obtained by mean shift clustering lies above the middle of the sign, leading to a possibly wrong position of the image patch for the following traffic sign detection. The fitted ellipse (sharp red circle) matches with the shape of the sign and allows therefore to extract the image patch for traffic sign recognition at the appropriate position

4 Results of traffic sign detection and recognition

The trained models for both traffic sign detection and recognition can obtain high overall accuracy values on the respective test set (97.2%, 93.4%). The approximately balanced number of samples in all classes ensures that the obtained overall accuracies do not result from models deciding always for the same class. It can be observed for the test set, that traffic signs with a very similar appearance (e.g. different speed limit signs) are the most challenging ones for traffic sign recognition. Further experiments using the skip architecture mentioned above in the neural networks have not shown to result in a higher overall accuracy.

As expected, traffic sign detection yields typically multiple detections for one traffic sign shown in a street scene image (Fig. 6). The larger the traffic sign in the image is, the higher the number of detections is. Mean shift clustering is able to reduce the detections to one cluster for large traffic signs (pink dot in Fig. 6). Critical attention has to be payed for remote traffic signs (i.e. small in the image), as observations (done using visualisations like in Fig. 6) have shown that adjacent small traffic signs might be grouped in one cluster by mistake. Similar observations can be made for ellipse fitting, which can be seen as robust for close traffic signs (i.e. large in the image).

Mean shift clustering and shape fitting as intermediate steps can be seen as useful to reduce the number of false positives resulting from traffic sign detection. Applied to the street scene images of the GTSDB dataset, the number of false positives would be 20 % higher if these intermediate steps are skipped. Hereby, a detection is seen as true positive if its image patch is overlapping with more than 50 % of the ground truth bounding box of a traffic sign. The comparison is done using the number of false positives instead of the false positive rate, as the high number of image patches resulting from the sliding window approach leads to a high number of true negatives and would lead to very similar false positive rates, which could be misinterpreted. Further, using the intermediate steps leads to a slightly higher number of true positives detections, which could be drawn back that adjusting the position and size of the image patch of a single detection, resulting from the intermediate steps, leads to a greater overlap with the ground truth label. Consequently, the number of false negatives is slightly reduced.

A further experiment underlines the demand that the image patch of a detection has to be centred precisely on the traffic sign for traffic sign recognition: the overall accuracy of traffic sign recognition decreases by 20 percentage points, if a test set of image patches is used, which are not centred on a traffic sign (randomly shifted by up to ±50 % of the patch size), compared to using a set of patches, which are centred.

5 Conclusion

In this contribution, a method for traffic sign detection and recognition in street scene images using two separate convolutional neural networks has been proposed. Detection with different models for different groups of traffic signs (e.g. red prohibitive signs, blue mandatory signs) and recognition with one joint model for all traffic signs can achieve an overall accuracy of around 95 % in average on the tested set of image patches. Applying the traffic sign detector to image patches extracted from a street scene image using a sliding window approach yields typically multiple detections for the same traffic sign. Without clustering multiple detections to a single detection and adjusting the position of the image patch by fitting an ellipse (for circular traffic signs) to the shape of a traffic sign detection, the number of false positive detections is 20 % higher for the test images. Future extensions to the proposed method can be made by integrating a traffic sign tracker to reduce the computational effort fort detection by reducing the search space for the detector in subsequent images of an image series. Furthermore, a method to fit additional shapes (e.g. rectangle, triangle) to the image patches of positive detections could be integrated to make the method better usable for varying types of traffic signs.

6 References

AHN, S. J., WARNECKE, H.-J. & KOTOWSKI, R., 1999. Systematic geometric image measurement errors of circular object targets: mathematical formulation and correction. Photogrammetric Record, **16**(93), 485-502.

COMASCHI, F., STUIJK, S., BASTEN, T. & CORPORAAL, H., 2013. RASW: a Run-time Adaptive Sliding Window to Improve Viola-Jones Object Detection. 2013 Seventh International Conference on Distributed Smart Cameras (ICDSC), 1-6.

CORDTS, M., OMRAN, M., RAMOS, S., REHFELD, T., ENZWEILER, M., BENENSON, R., FRANKE, U., ROTH, S. & SCHIELE, B., 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3213-3223.

ETKBMW.COM, 2018. Distance Systems, Cruise Control BMW G11 sedan 57553. Website, https://www.etkbmw.com/bmw/EN/search/selectCar/G11/Lim/BMW+750i/ECE/66. Accessed 2018-01-16.

EURO NCAP, 2013: Speed Assistance Systems. Website, https://www.euroncap.com/en/vehicle-safety/the-ratings-explained/safety-assist/speed-assistance/. Accessed 2018-01-16.

FUKUNAGA, K. & HOSTETLER, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory, **21**(1), 32-40.

HANEL, A. & STILLA, U., 2018. Iterative calibration of a vehicle camera using traffic signs detected by a convolutional neural network. Proceedings of the 4th International Conference on Vehicle Technology and Intelligent Transport Systems. Paper accepted; to be published

LARSSON, F. & FELSBERG, M., 2011. Using Fourier Descriptors and Spatial Models for Traffic Sign Recognition. Image Analysis: Proceedings of 17th Scandinavian Conference, SCIA 2011, Anders, H. & Kahl, F. (eds.), Springer Berlin / Heidelberg, 238-249.

REDMON, J., DIVVALA, S., GIRSHICK, R. & FARHADI, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779-788.

SERMANET, P. & LECUN, Y., 2011. Traffic sign recognition with multi-scale Convolutional Networks. The 2011 International Joint Conference on Neural Networks, 2809-2813.

STALLKAMP, J., SCHLIPSING, M., SALMEN, J. & IGEL, C., 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural Networks **32** (Supplement C), 323-332.

WU, Y., LIU, Y., LI, J., LIU, H. & HU, X., 2013. Traffic sign detection based on convolutional neural networks. The International Joint Conference on Neural Networks (IJCNN), 1-7.

ZHU, Z., LIANG, D., ZHANG, S., HUANG, X. & HU, S., 2016. Traffic-Sign Detection and Classification in the Wild. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2110-2118.