Neural Networks 32 (2012) 323–332

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2012 Special Issue

Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition

J. Stallkamp [a,*], M. Schlipsing [a], J. Salmen [a], C. Igel [b]

[a] Institut für Neuroinformatik, Ruhr-Universität Bochum, Universitätsstraße 150, 44780 Bochum, Germany

[b] Department of Computer Science, University of Copenhagen, Universitetsparken 1, 2100 Copenhagen, Denmark

ARTICLE INFO

Keywords:
Traffic sign recognition
Machine learning
Convolutional neural networks
Benchmarking

ABSTRACT

Traffic signs are characterized by a wide variability in their visual appearance in real-world environments. For example, changes of illumination, varying weather conditions and partial occlusions impact the perception of road signs. In practice, a large number of different sign classes needs to be recognized with very high accuracy. Traffic signs have been designed to be easily readable for humans, who perform very well at this task. For computer systems, however, classifying traffic signs still seems to pose a challenging pattern recognition problem. Both image processing and machine learning algorithms are continuously refined to improve on this task. But little systematic comparison of such systems exist. What is the status quo? Do today's algorithms reach human performance? For assessing the performance of state-of-the-art machine learning algorithms, we present a publicly available traffic sign dataset with more than 50,000 images of German road signs in 43 classes. The data was considered in the second stage of the German Traffic Sign Recognition Benchmark held at IJCNN 2011. The results of this competition are reported and the best-performing algorithms are briefly described. Convolutional neural networks (CNNs) showed particularly high classification accuracies in the competition. We measured the performance of human subjects on the same data—and the CNNs outperformed the human test persons.

1. Introduction

Traffic sign recognition is a multi-category classification problem with unbalanced class frequencies. It is a challenging real-world computer vision problem of high practical relevance, which has been a research topic for several decades. Many studies have been published on this subject and multiple systems, which often restrict themselves to a subset of relevant signs, are already commercially available in new high- and mid-range vehicles. Nevertheless, there has been little systematic unbiased comparison of approaches and comprehensive benchmark datasets are not publicly available.

Road signs are designed to be easily detected and recognized by human drivers. They follow clear design principles using color, shape, icons and text. These allow for a wide range of variations between classes. Signs with the same general meaning, such as the various speed limits, have a common general appearance, leading to subsets of traffic signs that are very similar to each other. Illumination changes, partial occlusions, rotations, and weather conditions further increase the range of variations in visual appearance a classifier has to cope with.

Humans are capable of recognizing the large variety of existing road signs in most situations with near-perfect accuracy. This does not only apply to real-world driving, where rich context information and multiple views of a single traffic sign are available, but also to the recognition from individual, clipped images.

In this paper, we compare the traffic sign recognition performance of humans to that of state-of-the-art machine learning algorithms. These results were generated in the context of the second stage of the *German Traffic Sign Recognition Benchmark* (GTSRB) held at IJCNN 2011. We present the extended GTSRB dataset with 51,840 images of German road signs in 43 classes. A website with a public leaderboard was set up and will be permanently available for submission of new results. Details about the competition design and analysis of the results of the first stage are described by Stallkamp, Schlipsing, Salmen, and Igel (2011).

The paper is organized as follows: Section 2 presents related work. Section 3 provides details about the benchmark dataset. Section 4 explains how the human traffic sign recognition performance is determined, whereas the benchmarked machine learning algorithms are presented in Section 5. The evaluation procedure is described in Section 6, together with the associated public leaderboard. Benchmarking results are reported and discussed in Section 7, before conclusions are drawn in Section 8.

* Corresponding author. Tel.: +49 234 3225566; fax: +49 234 3214210.
E-mail addresses: johannes.stallkamp@ini.rub.de (J. Stallkamp), marc.schlipsing@ini.rub.de (M. Schlipsing), jan.salmen@ini.rub.de (J. Salmen), igel@diku.dk (C. Igel).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.02.016

2. Related work

It is difficult to compare the published work on traffic sign recognition. Studies are based on different data and either consider the complete task chain of detection, classification and tracking or focus on the classification part only. Some articles concentrate on subclasses of signs, for example on speed limit signs and digit recognition.

Bahlmann, Zhu, Ramesh, Pellkofer, and Koehler (2005) present a holistic system covering all three processing steps. The classifier itself is claimed to operate with a correct classification rate of 94% on images from 23 classes. Training was conducted on 4000 traffic sign images featuring an unbalanced class frequency of 30–600 examples. The individual performance of the classification component is evaluated on a test set of 1700 samples.

Moutarde, Bargeton, Herbin, and Chanussot (2007) present a system for recognition of European and US speed limit signs. Their approach is based on single digit recognition using a neural network. Including detection and tracking, the proposed system obtains a performance of 89% for US and 90% for European speed limits, respectively, on 281 traffic signs. Individual classification results are not provided.

Another traffic sign detection framework is presented by Ruta, Li, and Liu (2010). The overall system including detection and classification of 48 different signs achieves a performance of 85.3% while obtaining classification error rates below 9%.

Broggi, Cerri, Medici, Porta, and Ghisio (2007) apply multiple neural networks to classify different traffic signs. In order to choose the appropriate network, shape and color information from the detection stage is used. The authors only provide qualitative classification results.

In the work by Keller, Sprunk, Bahlmann, Giebel, and Baratoff (2008), a number-based speed limit classifier is trained on 2880 images. It achieves a correct classification rate of 92.4% on 1233 images. However, it is not clear whether images of the same traffic sign instance are shared between sets.

Gao, Podladchikova, Shaposhnikov, Hong, and Shevtsova (2006) propose a system based on color features inspired by human vision. They report recognition rates up to 95% on 98 British traffic sign images.

Various approaches are compared on a dataset containing 1300 preprocessed examples from 6 classes (5 speed limits and 1 noise class) by Muhammad, Lavesson, Davidsson, and Nilsson (2009). The best classification performance observed was 97%.

In the study by Maldonado Bascón, Acevedo Rodríguez, Lafuente Arroyo, Caballero, and López-Ferreras (2010), a classification performance of 95.5% is achieved using support vector machines. The database comprises ∼36,000 Spanish traffic sign samples of 193 sign classes. However, it is not clear whether the training and test sets can be assumed to be independent, as the random split only took care of maintaining the distribution of traffic sign classes (see Section 3). To our knowledge, this database is not publicly available.

Obviously, the results reported above are not comparable, as all systems are evaluated on proprietary data, most of which is not publicly available. Therefore, we present a freely available, extensive traffic sign data set to allow unbiased comparison of traffic sign recognition approaches.

3. Dataset

This section describes our publicly available benchmark dataset. We explain the process of data collection and the provided data representation.

3.1. Data collection

The dataset was created from approx. 10 h of video that were recorded while driving on different road types in Germany during daytime. The sequences were recorded in March, October and November 2010. For data collection, a *Prosilica GC 1380CH* camera was used with automatic exposure control and a frame rate of 25 fps. The camera images, from which the traffic sign images are extracted, have a resolution of 1360 × 1024 pixels. The video sequences are stored in a raw *Bayer*-pattern format (Bayer, 1975).

Data collection, annotation and image extraction was performed using the *NISYS Advanced Development and Analysis Framework (ADAF)*,[1] an easily extensible, module-based software system (see Fig. 1).

We will use the term *traffic sign instance* to refer to a physical real-world traffic sign in order to discriminate against *traffic sign images* which are captured when passing the traffic sign by car. The sequence of images originating from one traffic sign instance will be referred to as a *track*. Each instance is unique. In other words, the dataset only contains a single track for each physical traffic sign.

3.2. Data organization

From 144,769 labelled traffic sign images of 2416 traffic sign instances in 70 classes, the GTSRB dataset was compiled according to the following criteria:

1. Discard tracks with less than 30 images.
2. Discard classes with less than 9 tracks.
3. For the remaining tracks: If the track contains more than 30 images, equidistantly sample 30 images.

Step 3 was performed for two reasons. First of all, the car passes different traffic sign instances with different velocities, depending on sign position and the overall traffic situation. In the recording, this leads to different numbers of traffic sign images per track (approximately 5–250 images per track). Consecutive images of a traffic sign that was passed with low velocity are very similar to each other. They do not contribute to the diversity of the dataset. On the contrary, they cause an undesired imbalance of dependent images. Since the different velocities are not uniformly distributed over all traffic sign types, this would strongly favour image classes that are present in low-speed traffic (*Stop*, *Yield-right-of-way*, low speed limits).

Secondly, the question arises why to keep multiple images per track at all. Although consecutive images in long tracks are nearly identical, the visual appearance of a traffic sign can vary significantly over the complete track, as can be seen in Fig. 2. Traffic signs at high distance result in low resolution while closer ones are prone to motion blur. The illumination may change, and the motion of the car affects the perspective with respect to occlusions and background. Selecting a fixed number of images per traffic sign both increases the diversity of the dataset in terms of the variations mentioned above and avoids an undesired imbalance caused by large numbers of nearly identical images.

The selection procedure outlined above reduced the number to 51,840 images of the 43 classes that are shown in Fig. 3. The relative class frequencies of the classes are shown in Fig. 4.

The set contains images of more than 1700 traffic sign instances. The size of the traffic signs varies between 15 × 15 and 222 × 193 pixels. The images contain 10% margin (at least 5 pixels) around the traffic sign to allow for the usage of edge detectors. The original size and location of the traffic sign within the image (region of interest, ROI) is preserved in the provided annotations. The images are not necessarily square. Fig. 5 shows the distribution of traffic sign sizes, taking into account the larger of both dimensions of the traffic sign ROI.

[1] http://www.nisys.de.

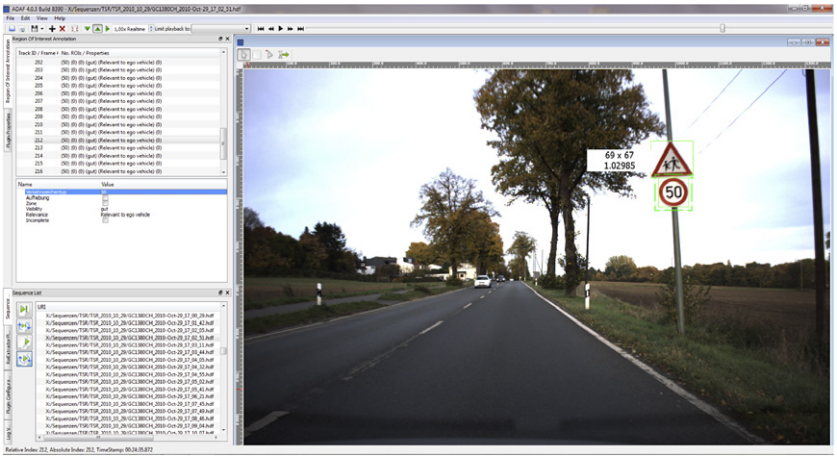


Fig. 1. Screenshot of the software used for the manual annotation. We made use of the NISYS Advanced Development and Analysis Framework (ADAF).

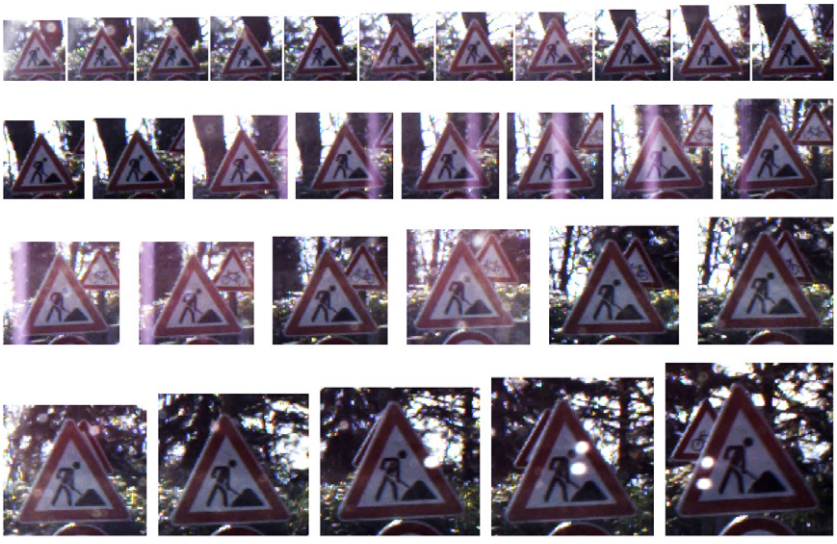


Fig. 2. A traffic sign *track*, which contains *traffic sign images* captured when passing a particular *traffic sign instance*.

Fig. 3. Random representatives of the 43 traffic sign classes in the GTSRB dataset.

The GTSRB dataset was split into three subsets according to Fig. 6. We applied stratified sampling. The split was performed at random, but taking into account class and track membership. This makes sure that (a) the overall class distribution is preserved for each individual set and that (b) all images of one traffic sign instance are assigned to the same set, as otherwise the datasets could not be considered stochastically independent.

The main split separates the data into the *full training set* and the *test set*. The training set is ordered by class. Furthermore, the images are grouped by tracks to preserve temporal information,

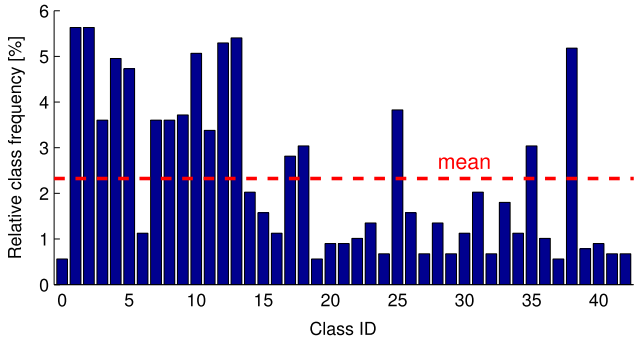


Fig. 4. Relative class frequencies in the dataset. The class ID results from enumerating the classes in Fig. 3 from top-left to bottom-right.

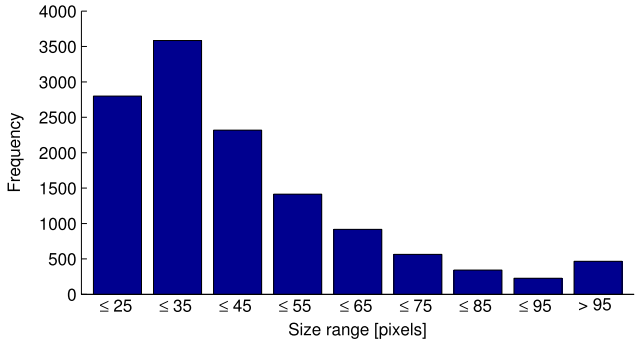


Fig. 5. Distribution of traffic sign sizes (in pixels).

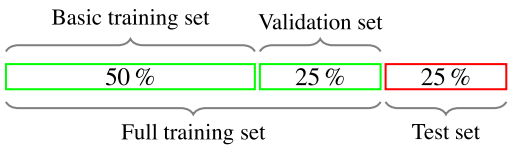


Fig. 6. For the two stages of the competition, the data was split into three sets.

which may be exploited by algorithms that are capable of using privileged information (Vapnik & Vashist, 2009). It can be used for final training of the classifier after all necessary design decisions were made or for training of parameter-free classifiers.

For the *test set*, in contrast to the training set, temporal information is not available. It is consecutively numbered and shuffled to prevent deduction of class membership from other images of the same track.

For the first stage of the GTSRB (see Section 5 and Stallkamp et al., 2011), the full training set was partitioned into two sets. The *validation set* is a subset of the full training set and is still provided for convenience. It is generated according to the aforementioned criteria and, thus, ensures a consistent class distribution and clean separation from the other sets. It allows for classifier selection, parameter search and optimization. Data in the validation set is available in two different configurations: (a) shuffled like the test set, which allows a fixed system setup for training and testing, and (b) appended to the *basic training set* – sorted by class and grouped by track – as part of the *full training set*. The validation set played the role of the test set in the online competition (see Stallkamp et al., 2011 and Section 5).

3.3. Data representation

To allow participants without image processing background to benchmark their machine learning approaches on the data, all sets are provided in different representations. The following pre-calculated features are included:

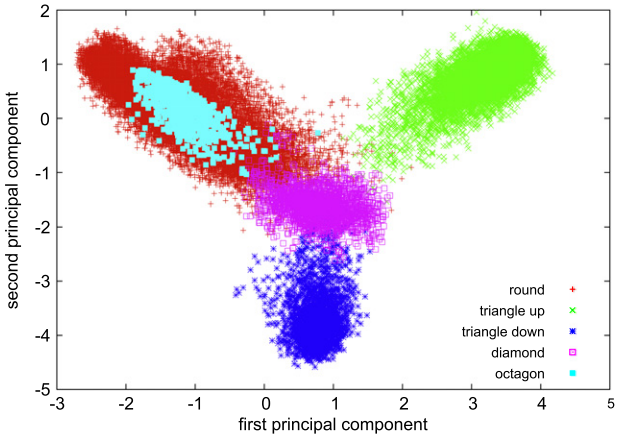


Fig. 7. The "HOG1" training data projected on its first two principal components.

Fig. 8. Haar features types used to generate one of the representations provided by the competition organizers.

3.3.1. Color images

Originally, the videos are recorded by a *Bayer* sensor array. All extracted traffic sign images are converted into *RGB* color images employing an edge-adaptive, constant-hue demosaicking method (Gunturk, Glotzbach, Altunbasak, Schafer, & Mersereau, 2005; Ramanath, Snyder, Bilbro, & Sander, 2002). The images are stored in *PPM* format alongside the corresponding annotations in a text file.

3.3.2. HOG features

Histograms of Oriented Gradient (HOG) descriptors have been proposed by Dalal and Triggs (2005) for pedestrian detection. Based on the gradients of color images, different weighted and normalized histograms are calculated: first for small non-overlapping *cells* of multiple pixels that cover the whole image and then for larger overlapping *blocks* that integrate over multiple cells.

We provided three sets of features from differently configured HOG descriptors, which we expected to perform well when used for classification. To compute HOG features, all images were scaled to a size of 40 × 40 pixels. For sets 1 and 3 the sign of the gradient response was ignored. Sets 1 and 2 use cells of size 5 × 5 pixels, a block size of 2 × 2 cells and an orientation resolution of 8, resulting in feature vectors of length 1568. In contrast, for "HOG 3", cells of size 4 × 4 pixels and 9 orientations resulted in 2916 features.

HOG descriptors provide a good representation of the traffic signs. As can be seen in Fig. 7, the first two principal components already provide a clear and meaningful separation of different sign shapes (e.g., the diamond shaped signs are located between the upwards and downwards pointing triangular signs).

3.3.3. Haar-like features

The popularity of Haar features is mainly due to the efficient computation using the *integral image* proposed by Viola and Jones (2001) and their outstanding performance in real-time object detection employing a cascade of weak classifiers.

Just as for the HOG features, images were rescaled to 40 × 40 pixels. In order to compute Haar features, they were converted to grayscale after rescaling. We computed five different types (see Fig. 8) in different sizes to a total of 11,584 features per image. While one would usually apply feature selection (Salmen, Schlipsing, & Igel, 2010) we provide all Haar-feature responses in the set.

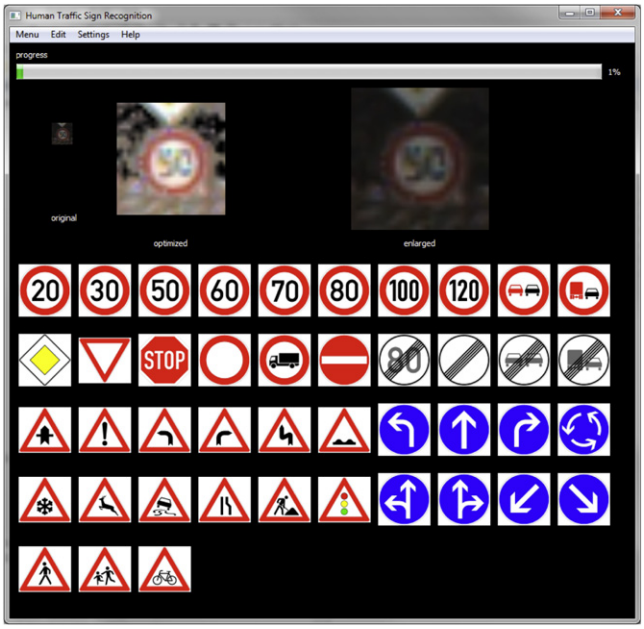


Fig. 9. User interface of the human performance application.

3.3.4. Color histograms

This set of features was provided to complement the gradient-based feature sets with color information. It contains a global histogram of the hue values in HSV color space, resulting in 256 features per image.

4. Human performance

Traffic signs are designed to be easily distinguishable and readable by humans. Once spotted, recognition of the majority of traffic signs is not a challenging problem for them. Although real-life traffic provides rich context, it is not required for the task of pure classification. Humans are well capable of recognizing the type of a traffic sign from clipped images such as in the GTSRB dataset (e.g., see Fig. 3).

In order to determine the human traffic sign recognition performance, two experiments were conducted. During these experiments, images were presented to the test person in three different versions (see Fig. 9): the original image, an enlarged version to improve readability of small images and a contrast-enhanced, enlarged version to improve readability of dark and low-contrast samples like the example in the Fig. 9. The test person assigned a class ID by clicking the corresponding button. Please note that this class ID assignment was for testing purposes only, not for generation of the ground-truth data, as this was done on the original camera images (see Section 3.1 and Fig. 1).

For the first experiment, the images in the test set were presented in chunks of 400 randomly chosen images each to 32 test persons. Over the complete course of the experiment, each image was presented exactly once for classification. This yielded an *average* traffic sign recognition performance over all subjects. This experiment was executed in analogy to the online competition (Stallkamp et al., 2011).

As shown in Fig. 10, there is some variance w.r.t. the individual performance. To some extent, this can be explained by the random selection of images that were presented to each of the subjects. Somebody with a lower performance might just have got more difficult images than somebody else with higher performance.

To eliminate this possibility, we set up another experiment to determine the traffic sign recognition performance of individual subjects on the full test set (12,630 images). As manual classification of this amount of data is a very tedious, time-consuming and concentration-demanding task, the experiment was limited to a single *well-performing* test person.

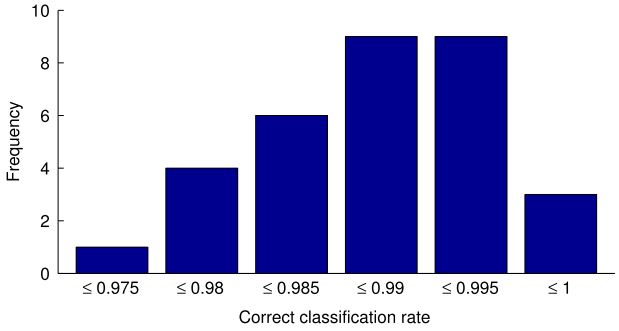


Fig. 10. Distribution of individual performance in the *average* human performance experiment.

To find a suitable candidate, we performed a *model selection* step, very much in the same sense as it is used when choosing or tuning a classifier for a problem. Eight test persons were confronted with a randomly selected, but fixed subset of 500 images of the validation set. The best-performing one was selected to classify the test set. In addition to selecting a candidate, the model selection step served as an initial training phase to get used to the sometimes unfamiliar appearance of traffic signs in the dataset. To reduce the negative impact of decreasing concentration on recognition performance, the experiment on the full test set was split into multiple sessions.

5. Benchmarked methods

This section describes the machine learning algorithms that were evaluated on the GTSRB dataset. This evaluation constituted the second stage of the IJCNN 2011 competition *The German Traffic Sign Recognition Benchmark* and was performed at the conference. The first stage of the competition – conducted online before the conference – attracted more than 20 teams from all around the world (Stallkamp et al., 2011). A wide range of state-of-the-art machine learning methods was employed, including (but not limited to) several kinds of neural networks, support vector machines, linear discriminant analysis, subspace analysis, ensemble classifiers, slow feature analysis, kd-trees, and random forests. The top teams were invited to the conference for a final competition session. However, participation was not limited to these teams. Any researcher or team could enter, regardless of their participation or performance in the first stage of competition. The second stage was set to reproduce or improve the results of the online stage and to prevent potential cheating.

In addition to a baseline algorithm, we present the approaches of the three best-performing teams.

5.1. Baseline: LDA

As a baseline for comparison, we provide results of a linear classifier trained by linear discriminant analysis (LDA). Linear discriminant analysis is based on a *maximum a posteriori* estimate of the class membership. The classification rule is derived under the assumption that the class densities are multi-variate Gaussians having a common covariance matrix. Linear discrimination using LDA gives surprisingly good results in practice despite its simplicity (Hastie, Tibshirani, & Friedman, 2001). The LDA was based on the implementation in the Shark Machine Learning Library (Igel, Glasmachers, & Heidrich-Meisner, 2008), which is publicly available.[2]

[2] http://shark-project.sourceforge.net.

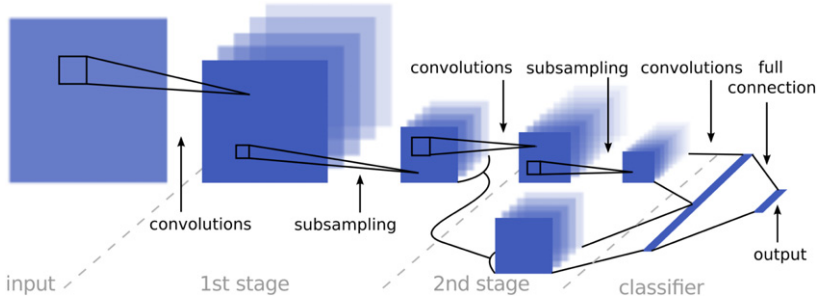


Fig. 11. CNN architecture employed by Sermanet and LeCun (2011), who kindly provided this figure.

5.2. Team sermanet: multi-scale CNN

Sermanet and LeCun (2011) employed a multi-scale convolutional neural network (CNN or ConvNet). CNNs are biologically inspired multi-layer feed-forward networks that are able to learn task-specific invariant features in a hierarchical manner, as sketched in Fig. 11. The multiple feature extraction stages are trained using supervised learning. The raw images are used as input. Each feature extraction stage of the network consists of a convolutional layer, a non-linear transformation layer and a spatial pooling layer. The latter reduces the spatial resolution, which leads to improved robustness against small translations, similar to "complex cells" in the standard models of the visual cortex. In contrast to traditional CNNs, not only the output of the last stage but of all feature extraction stages are fed into the classifier. This results in a combination of different scales of the receptive field, providing both global and local features. Moreover, Sermanet and LeCun employed alternative non-linearities. They used a combination of a rectified sigmoid followed by subtractive and divisive local normalization, inspired by computational neuroscience models of vision (Lyu & Simoncelli, 2008; Pinto, Cox, & DiCarlo, 2008).

The input was scaled to a size of 32 × 32 pixels. Color information was discarded and the resulting grayscale images were contrast-normalized. To increase the robustness of the classifier, Sermanet and LeCun increased the training set size five-fold by perturbing the available samples with small, random changes of translation, rotation and scale.

5.3. Team IDSIA: committee of CNNs

Team *IDSIA* used a committee of CNNs in the form of a multi-column deep neural network (MCDNN). It is based on a flexible, high-performance GPU implementation. The approach in Ciresan, Meier, Masci, and Schmidhuber (2011) won the first stage of the GTSRB competition by using a committee of CNNs trained on raw image pixels and multi-layer perceptrons (MLP) trained on the three provided HOG feature sets. For the second and final competition stage, for which results are presented in this paper, the authors dropped the MLPs. In turn, they increased the number of DNNs, because MCDNN with more columns showed improved performance. The details on the architecture for one DNN is shown in Table 1.

In contrast to team *Sermanet* (see 5.2), team *IDSIA* only uses the central ROI containing the traffic sign and ignores the margin. This region is scaled to a size of 48 × 48 pixels. In comparison to their approach for the online competition, the authors improved the preprocessing of the data by using four image adjustments methods. Histogram stretching increases image contrast by remapping pixel intensities to use the full range of available values. Histogram equalization transforms pixel intensities so that the histogram of the resulting image is approximately uniform. Adaptive histogram equalization applies the same principle, but to non-overlapping tiles rather than the full image. Contrast normalization enhances edges by filtering the image with a difference of Gaussians. The latter was inspired by the approach of team *Sermanet*. Each preprocessing step was applied individually to the training data, resulting in a five-fold increase of the number of training samples. The generalization of the individual networks is further increased by random perturbations of the training data in terms of translation, rotation and scale. However, in contrast to team *Sermanet*, these distortions are computed on-the-fly every time an image is passed through the network during training. Thus, every image is distorted differently in each epoch. The training of each DNN requires about 25 epochs and takes about 2 h. This leads to a total training time of approximately 50 h for MCDNN.

Table 1
8-layer DNN architecture used by team IDSIA.

Layer	Type	# Maps	Neurons/map	Kernel
0	Input	3	48 × 48	
1	Convolutional	100	42 × 42	7 × 7
2	Max pooling	100	21 × 21	2 × 2
3	Convolutional	150	18 × 18	4 × 4
4	Max pooling	150	9 × 9	2 × 2
5	Convolutional	250	6 × 6	4 × 4
6	Max pooling	250	3 × 3	2 × 2
7	Fully connected		300	1 × 1
8	Fully connected		43	1 × 1

5.4. Team CAOR: random forests

The competition entry of team *CAOR* is based on a Random Forest of 500 trees. A Random Forest is an ensemble classifier that is based on a set of non-pruned random decision trees (Breiman, 2001). Each decision tree is built on a randomly chosen subset of the training data. The remaining data is used to estimate the classification error. In each node of a tree, a small, randomly chosen subset of features is selected and the best split of the data is determined based on this selection. For classification, a sample is passed through all decision trees. The outcome of the Random Forest is a majority vote over all trees. Team *CAOR* used the official HOG 2 dataset. More details on this approach are reported by Zaklouta, Stanciulescu, and Hamdoun (2011).

6. Evaluation procedure

Participating algorithms need to classify the single images of the GTSRB test set. For model selection and training of the final classifier, the basic training set and the validation set (cf. Section 3) can be used either independently or combined (full training set).

Here, we explain how the performance of the algorithms is assessed and introduce the benchmark website, featuring a public leaderboard and detailed result analysis.

6.1. Performance metric

The performance is evaluated based on the 0/1 loss, that is, by basically counting the number of misclassifications. Therefore, we are able to rank algorithms based on their empirical *correct classification rate* (CCR).

The loss is chosen equal for all misclassifications, although the test set is strongly unbalanced w.r.t. the number of samples per class. This accounts for the fact that every sign is equally important, independent of variable frequencies of appearance. Nevertheless, the performance for the different subsets is also considered separately (see Section 7.4).

6.2. Public leaderboard

In addition to the benchmark dataset itself, we provide an evaluation website[3] featuring a public leaderboard. It was inspired by a similar website for comparison of stereo vision algorithms[4] established by Scharstein and Szeliski (2002). Fig. 12 shows a screenshot of the GTSRB submission website.

Our benchmark website will remain permanently open for submissions. It allows participants to upload result files (in a simple CSV format) and get immediate feedback about their performance. The results can be made publicly visible as soon as publication details are provided. Approaches are ranked based on their performance on the whole test dataset. Nevertheless, we allow re-sorting based on subset evaluation.

The website provides a more detailed result analysis, for instance online computation of the confusion matrix and a list of all misclassified images. For even more detailed offline analysis, an open-source software application can be downloaded that additionally enables participants to compare multiple approaches.

We encourage researchers to continue submitting their results. While different machine learning algorithms already have been shown to achieve very high performance, there is a particular interest in having more real-time capable methods or approaches focusing on difficult subsets.

7. Results and discussion

We report the classification performance of the three best-performing machine learning approaches complemented with the results of the baseline algorithm as described in Section 5. Furthermore, we present the results of the experiments on human traffic sign recognition performance (see Section 4). The results that are reported in this section are summarized in Table 2.

7.1. Human performance

For a human observer, the images in the dataset vary strongly in terms of quality and readability. This is, to a large extent, caused by visual artifacts – such as low resolution, low contrast, motion blur, or reflections – which originate from the data acquisition process and hardware. Although the machine learning algorithms have to deal with these issues as well, the visual appearance of traffic signs in deficient images can be very unfamiliar to human observers compared to the traffic signs they encounter in reality.

As noted in Section 4, the first experiment on human performance yields an *average* traffic sign recognition rate over all subjects. The distribution of individual classification performances of the 32 test persons is shown in Fig. 10. However, this does not give a clear picture of human traffic sign recognition performance, as the individual image sets that were presented to the test subjects could vary significantly in difficulty due to the aforementioned reasons. Although the test application is designed to improve readability of low-quality images and, thus, reduce the impact of this variation of difficulty, it cannot resolve the issues completely. Therefore, the variations of individual performance are caused both by the varying difficulty of the selected images and by the differing ability of the subjects to cope with these issues and to actually recognize the traffic signs. The model selection step of the second human performance experiment prevents the former issue by using a random *but fixed* dataset. Thus, the varying performance in this experiment is due to the individual ability of the test persons. As can be seen in Table 2, the single best test person performs significantly better (McNemar's test, $p < 0.001$) than the average, reaching an accuracy of 99.22%. Therefore, future references in this section refer to the human performance of the single best individual.

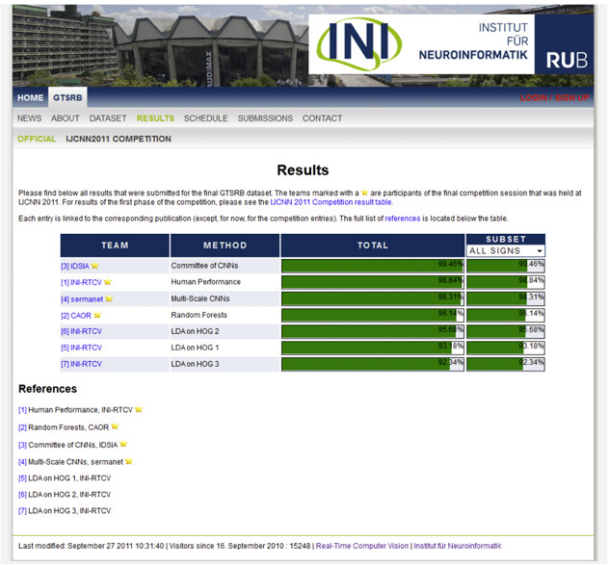


Fig. 12. The GTSRB submission website, which is open for new contributions.

Table 2
Result overview for the final stage of the GTSRB.

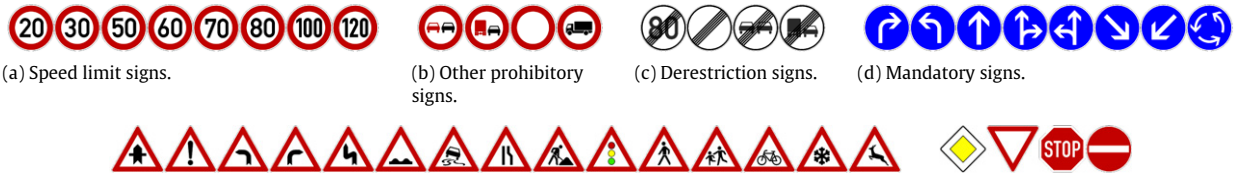
CCR (%)	Team	Method
99.46	IDSIA	Committee of CNNs
99.22	INI-RTCV	Human (best individual)
98.84	INI-RTCV	Human (average)
98.31	Sermanet	Multi-scale CNN
96.14	CAOR	Random forests
95.68	INI-RTCV	LDA (HOG 2)
93.18	INI-RTCV	LDA (HOG 1)
92.34	INI-RTCV	LDA (HOG 3)

7.2. Machine learning algorithms

As can be seen in Table 2, most of the machine learning algorithms achieved a correct recognition rate of more than 95%, with the committee of CNNs reaching near-perfect accuracy, outperforming the human test persons.

From an application point of view, processing time and resource requirements are important aspects when choosing a classifier. In this context, it is notable how well LDA – a very simple and computationally cheap classifier – performs in comparison to the more complex approaches. Especially the convolutional networks are computationally demanding, both during training and testing. Not surprisingly, the performance of LDA was considerably dependent on the feature representation. In the following, we

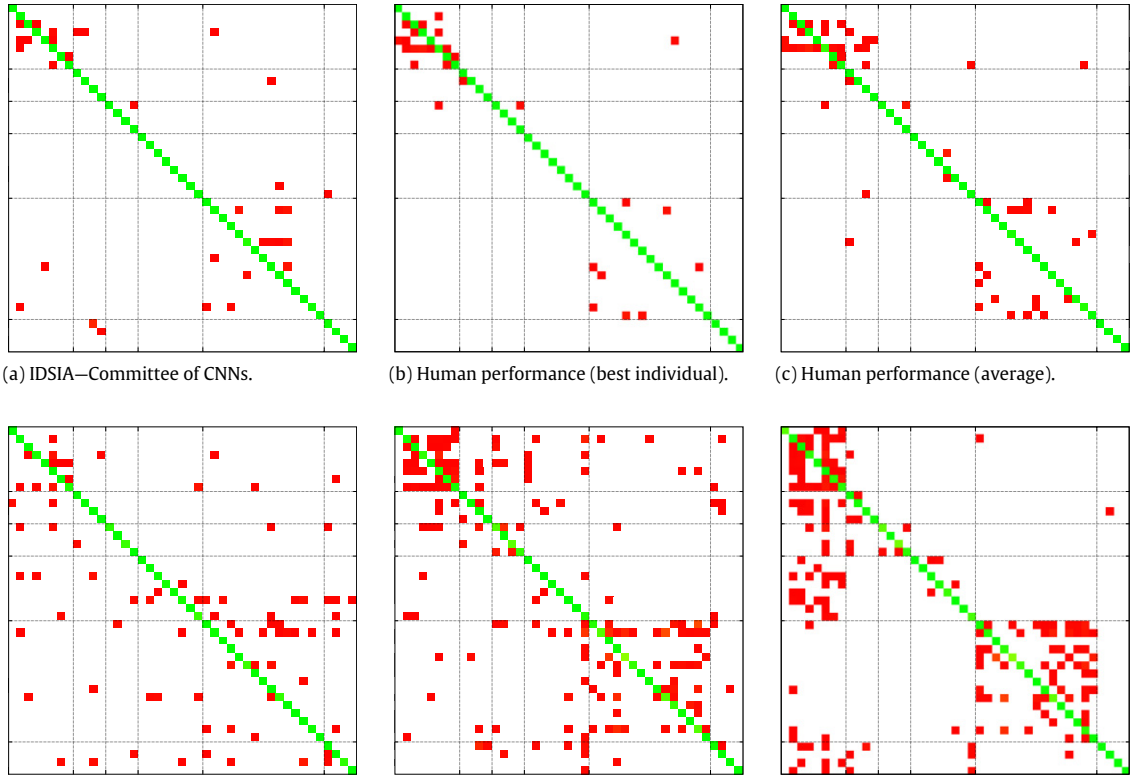
[3] http://benchmark.ini.rub.de.
[4] http://vision.middlebury.edu/stereo.



(a) Speed limit signs. (b) Other prohibitory signs. (c) Derestriction signs. (d) Mandatory signs.

(e) Danger signs. (f) Unique signs.

Fig. 13. Subsets of traffic signs.



(a) IDSIA—Committee of CNNs. (b) Human performance (best individual). (c) Human performance (average).

(d) Sermanet—multi-scale CNN. (e) CAOR—random forests. (f) LDA.

Fig. 14. Confusion matrices. The grid lines separate the traffic sign subsets defined in Fig. 13. The encoded values are normalized per class and in the range [0, 1].

Table 3
Individual results for subsets of traffic signs. Bold type denotes the best result(s) per subset.

	Speed limits	Other prohibitions	Derestriction	Mandatory	Danger	Unique
Committee of CNNs	**99.47**	**99.93**	**99.72**	99.89	99.07	99.22
Human (best individual)	98.32	99.87	98.89	**100.00**	**99.21**	**100.00**
Human (average)	97.63	**99.93**	98.89	99.72	98.67	**100.00**
Multi-scale CNN	98.61	99.87	94.44	97.18	98.03	98.63
Random forests (HOG 2)	95.95	99.13	87.50	99.27	92.08	98.73
LDA (HOG 2)	95.37	96.80	85.83	97.18	93.73	98.63

just refer to the best LDA results achieved with the HOG 2 representation. The performance results of the machine learning algorithms are all significantly different from each other. With exception of the comparison of Random Forests and LDA ($p = 0.00865$), all pairwise p-values are smaller than 10^{-10}. The values were calculated with McNemar's test for paired samples.[5]

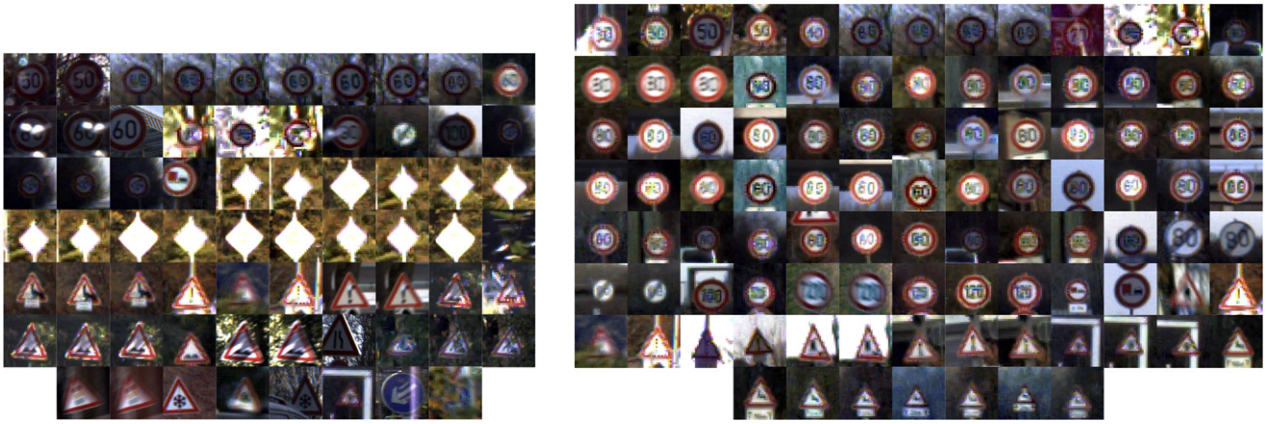
7.3. Man vs. computer

Both the best human individual and the best machine learning algorithm achieve a very high classification accuracy. The Committee of CNNs performs significantly better than the best human individual (McNemar's test, $p = 0.01366$). However, even without taking into account that the experimental setup for the human performance was unfamiliar for the test subjects and did not reflect real-life traffic scenarios, it needs to be noted that the best human test person significantly outperformed all other machine learning algorithms in this comparison. All pairwise p-values, as calculated with McNemar's test, are smaller than 10^{-10}.

7.4. Subsets

In order to gain a deeper insight into the results, we split the dataset into groups of similar traffic sign classes, as shown in Fig. 13. The individual results per approach and subset are listed in

[5] We provide and discuss p-values instead of confidence levels to show that correcting for multiple testing still leads to significant results.



(a) Committee of CNNs. (b) Human performance (best individual).

Fig. 15. Incorrectly classified images.

Table 3. A more detailed view is provided by the confusion matrices for the different approaches in Fig. 14. The classes are ordered by subsets as defined in Fig. 13(a)–(f), from left-to-right and top-to-bottom respectively. Rows denote the true class, columns the assigned class. The subsets are separated by the grey lines. The confusion matrices show the distribution of error over the different classes.

Common to all approaches except the multi-scale CNN, although to different extents, is a clustering in two areas: in the top-left corner, which corresponds to the subset of *speed limit* signs (see Fig. 13(a)), and in the large area in the lower right (second last row/column), which corresponds to the subset of triangular *danger* signs (see Fig. 13(e)). As can be seen in Fig. 14, the signs in these subsets are mostly mistaken for signs in the same subset. So the general shape is matched correctly, but the contained number or icon cannot be discriminated. If a traffic sign provided less-detailed content, like the blue *mandatory* signs (see Fig. 13(d)), or if the sign has a very distinct shape, such as the *unique* signs (see Fig. 13(f)), the recognition rate is usually above average, with humans even achieving perfect accuracy.

The HOG-based LDA is able to discriminate the round signs from the triangular ones. However, it easily confuses all round signs (and some of the unique signs as well) for *speed limits*. This is caused by the strongly imbalanced dataset, in which a third of all signs belong to this subset.

Although similar in overall performance, the Random Forest approach is not affected by this imbalance. Each decision tree in the forest is trained on a different, random sample of the training data. Therefore, the class distribution in this sample can be very different from the overall dataset.

7.5. Incorrectly classified images

Visual inspection of errors allows one to better understand why a certain approach failed at correct classification. Fig. 15 shows the images that were incorrectly classified by the best machine learning approach and by the best individual in the human performance experiment. For presentation purposes, all images were contrast-enhanced and scaled to a fixed size.

It is notable that a large part of the error of the committee of CNNs is caused by a single traffic sign instance, a diamond-shaped *right-of-way* sign. It accounts for more than 15% of the total error. However, not *all* images of this traffic sign track were misclassified, but only half of them. In fact, the committee misclassified those images in this track that were so overexposed that the yellow center is mostly lost. For humans, this sign class generally poses no problem due to its unique shape.

Furthermore, the algorithm misclassified a few images due to occlusion (such as reflections and graffiti) and two images due to inaccurate annotation that resulted in a non-centered view of the traffic sign. These images are easily classified by humans.

In contrast, the most difficult class for humans are *speed limit* signs, especially at low resolution, which impairs discrimination of single digits and, thus, correct recognition. More than 70% percent of the error can be accounted to this subset of traffic signs. Misclassification of *danger* signs causes the major part of the remaining error for the same reasons. Typical examples for confusion are caused by similar structures, for example the exclamation mark (general *danger* sign) being confused for the *traffic light* sign and vice versa (second and ninth traffic sign in Fig. 13(e)), or the *curvy road* sign being confused with *crossing deer* (fifth and last traffic sign in Fig. 13(e)), which both show a diagonal distibution of black pixels in the icon area.

7.6. Image size

As shown in Fig. 5, the images in the dataset vary strongly in size. Smaller images provide lower resolution by definition, whereas the very large images, i.e., the ones of traffic signs in close proximity to the ego vehicle, often show blurring or ghost images (showing the sign twice, blurry and slightly shifted) due to the larger relative motion in the image plane. Fig. 16 shows the classification performance of all presented approaches in dependency of the image size. It is not surprising that, for all approaches, the recognition rate is the lowest for the smallest images. The low resolution strongly impairs discriminability of fine details such as the single digits on *speed limit* signs or the icons on *danger* signs. The human performance continuously increases with increasing image size, reaching perfect accuracy for images larger than 45 pixels (in the larger of both dimensions) for the best individual and for images larger than 75 pixels in the average case. The algorithmic approaches, however, show reduced performance for very close images. Possible reasons are the strong motion blur or the presence of ghost images, such as in the lower left images in Fig. 15(a).

This reduction of performance is strongest for Random Forests and LDA, which generally show a very similar performance when different image sizes are considered. In addition, both approaches show a major impact on recognition performance for very small images. Contrary to expectation, the smallest error does not occur for mid-size images, which often are of good quality in terms of resolution and blurring. As the number of images per size level is strongly decreasing with increasing image size (see Fig. 5), the sensitivity to single misclassified tracks (or large parts thereof) increases and impairs performance.

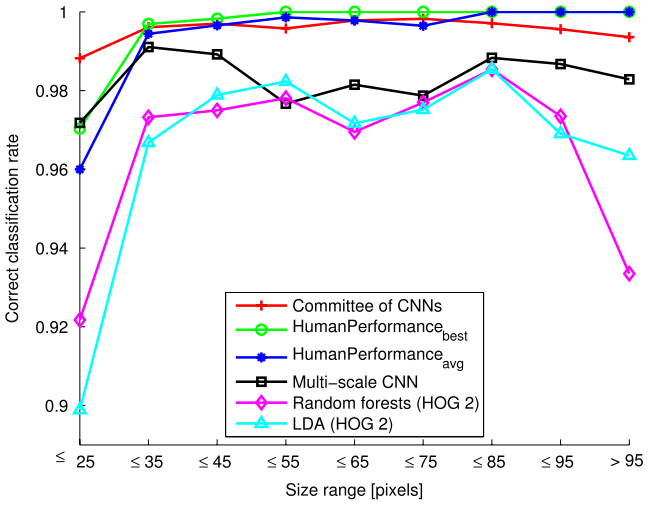


Fig. 16. Recognition performance depending on image size.

8. Conclusions

We presented a detailed comparison of the traffic sign recognition performance of state-of-the-art machine learning algorithms and humans. Although the best individual in the human performance experiment achieved a close-to-perfect accuracy of 99.22%, it was outperformed in this challenging task by the best-performing machine learning approach, a committee of convolutional neural networks, with 99.46% correct classification rate. In contrast to traditional computer vision, where hand-crafted features are common, convolutional neural networks are able to learn task-specific features from raw data. However, in return, "finding the optimal architecture of a ConvNet for a given task remains mainly empirical" (Sermanet & LeCun, 2011, Sec. II.B).

Moreover, convolutional neural networks are still computationally very demanding. Taking into account potential constraints on hardware capabilities and processing time, as they are common in the domain of driver assistance systems, it is striking to see how well linear discriminant analysis, a computationally cheap classifier, performs on this problem, reaching a correct recognition rate of 95.68%.

However, none of the machine learning approaches is able to handle input images of variable size and aspect ratio as present in the dataset. The usual approach is scaling of the images to a fixed size. This can cause problems when the aspect ratio is different between the original and target sizes. Furthermore, it discards information in larger images or introduces artifacts if very small images are strongly magnified. Humans are well capable of recognizing traffic signs of different size, even if viewed from sharp angles.

The public leaderboard on the competition website will be permanently open for submission and analysis of new results on the GTSRB dataset. For the future, we plan to add more benchmark tasks and data to the competition website. In particular, we are currently working on a benchmark dataset for the detection of traffic signs in full camera images.

Acknowledgments

We thank Lukas Caup, Sebastian Houben, Stefan Tenbült and Marc Tschentscher for their labelling support, Bastian Petzka for creating the competition website, NISYS GmbH for supplying the data collection and annotation software. We thank Fatin Zaklouta, Pierre Sermanet and Dan Ciresan for the valuable comments on their approaches. Furthermore, we want to thank all our test persons for the human performance experiment, especially Lisa Kalbitz, and all others that contributed to this competition. We acknowledge support from the German Federal Ministry of Education and Research within the National Network Computational Neuroscience, Bernstein Fokus: "Learning behavioral models: From human experiment to technical assistance", grant FKZ 01GQ0951.

References

Bahlmann, C., Zhu, Y., Ramesh, V., Pellkofer, M., & Koehler, T. (2005). A system for traffic sign detection, tracking, and recognition using color, shape, and motion information. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 255–260). IEEE Press.

Bayer, B. E. (1975). *US patent 3971065: color imaging array*. Eastman Kodak Company.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Broggi, A., Cerri, P., Medici, P., Porta, P. P., & Ghisio, G. (2007). Real time road signs recognition. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 981–986). IEEE Press.

Ciresan, D. C., Meier, U., Masci, J., & Schmidhuber, J. (2011). A committee of neural networks for traffic sign classification. In *Proceedings of the IEEE international joint conference on neural networks* (pp. 1918–1921). IEEE Press.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 886–893).

Gao, X. W., Podladchikova, L., Shaposhnikov, D., Hong, K., & Shevtsova, N. (2006). Recognition of traffic signs based on their colour and shape features extracted using human vision models. *Journal of Visual Communication and Image Representation*, *17*(4), 675–685.

Gunturk, B. K., Glotzbach, J., Altunbasak, Y., Schafer, R. W., & Mersereau, R. M. (2005). Demosaicking: color filter array interpolation. *IEEE Signal Processing Magazine*, *22*(1), 44–54.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag.

Igel, C., Glasmachers, T., & Heidrich-Meisner, V. (2008). Shark. *Journal of Machine Learning Research*, *9*, 993–996.

Keller, C. G., Sprunk, C., Bahlmann, C., Giebel, J., & Baratoff, G. (2008). Real-time recognition of US speed signs. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 518–523). IEEE Press.

Lyu, S., & Simoncelli, E. P. (2008). Nonlinear image representation using divisive normalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–8). IEEE Press.

Maldonado Bascón, S., Acevedo Rodríguez, J., Lafuente Arroyo, S., Caballero, A., & López-Ferreras, F. (2010). An optimization on pictogram identification for the road-sign recognition task using SVMs. *Computer Vision and Image Understanding*, *114*(3), 373–383.

Moutarde, F., Bargeton, A., Herbin, A., & Chanussot, A. (2007). Robust on-vehicle real-time visual detection of American and European speed limit signs with a modular traffic signs recognition system. In *Proceedings of the IEEE intelligent vehicles symposium* (pp. 1122–1126). IEEE Press.

Muhammad, A. S., Lavesson, N., Davidsson, P., & Nilsson, M. (2009). Analysis of speed sign classification algorithms using shape based segmentation of binary images. In *Proceedings of the international conference on computer analysis of images and patterns: Vol. 5702* (pp. 1220–1227). Springer-Verlag.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, *4*(1), e27.

Ramanath, R., Snyder, W. E., Bilbro, G. L., & Sander, W. A. (2002). Demosaicking methods for Bayer color arrays. *Journal of Electronic Imaging*, *11*, 306–315.

Ruta, A., Li, Y., & Liu, X. (2010). Real-time traffic sign recognition from video by class-specific discriminative features. *Pattern Recognition*, *43*(1), 416–430.

Salmen, J., Schlipsing, M., & Igel, C. (2010). Efficient update of the covariance matrix inverse in iterated linear discriminant analysis. *Pattern Recognition Letters*, *31*(1), 1903–1907.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, *47*, 7–42.

Sermanet, P., & LeCun, Y. (2011). Traffic sign recognition with multi-scale convolutional networks. In *Proceedings of the IEEE international joint conference on neural networks* (pp. 2809–2813). IEEE Press.

Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2011). The German traffic sign recognition benchmark: a multi-class classification competition. In *Proceedings of IEEE international joint conference on neural networks* (pp. 1453–1460). IEEE Press.

Vapnik, V., & Vashist, A. (2009). A new learning paradigm: learning using privileged information. *Neural Networks*, *22*(5–6), 544–557.

Viola, P., & Jones, M. (2001). Robust real-time object detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Zaklouta, F., Stanciulescu, B., & Hamdoun, O. (2011). Traffic sign classification using k-d trees and random forests. In *Proceedings of the IEEE international joint conference on neural networks* (pp. 2151–2155). IEEE Press.