Application of machine learning methods for traffic signs recognition

D V Filatov, K V Ignatev, A V Deviatkin and E V Serykh
V. Ulyanov (Lenin) St Petersburg State Electrotechnical University 'LETI' 5 Professora Popova Ul. St Petersburg 197376, Russian Federation

E-mail: evserykh@etu.ru

Abstract. This paper focuses on solving a relevant and pressing safety issue on intercity roads. Two approaches were considered for solving the problem of traffic signs recognition; the approaches involved neural networks to analyze images obtained from a camera in the real-time mode. The first approach is based on a sequential image processing. At the initial stage, with the help of color filters and morphological operations (dilatation and erosion), the area containing the traffic sign is located on the image, then the selected and scaled fragment of the image is analyzed using a feedforward neural network to determine the meaning of the found traffic sign. Learning of the neural network in this approach is carried out using a backpropagation method. The second approach involves convolution neural networks at both stages, i.e. when searching and selecting the area of the image containing the traffic sign, and when determining its meaning. Learning of the neural network in the second approach is carried out using the intersection over union function and a loss function. For neural networks to learn and the proposed algorithms to be tested, a series of videos from a dash cam were used that were shot under various weather and illumination conditions. As a result, the proposed approaches for traffic signs recognition were analyzed and compared by key indicators such as recognition rate percentage and the complexity of neural networks' learning process.

1. Introduction
Ensuring road safety is one of the most pressing issues in modern society. Safety issue is directly related to the 'human factor'. Driver and pedestrian inattention is one of the main causes of road traffic accidents. Among the key tasks for both active driver assistance systems (ADAS) and control systems for autonomous vehicles, the analysis of marking and traffic signs occupies an important place. Numerous computer vision systems have been developed recently for traffic signs analysis. But the characteristics of existing algorithms [1 – 5] (recognition accuracy, quantity of errors, robustness against atmospheric changes) are still not good enough to rule out a human operator.

Main tasks for computer vision systems used for the traffic signs analysis are detecting the area on a given frame that contains a road sign (localization task) and determining the meaning of the sign(s) on the selected area of the frame (classification task). The localization task can be solved in the following ways: using a deep neural network for predicting the location of the object's describing rectangle [6]; using methods for determining the parts of objects and relations between them [7 - 9].

Classification can be performed using the following methods: sliding window method [10 – 16]; gradient boosting over decision trees [17].
In the field of the development of intelligent systems for signs recognition, there are works focusing on both feedforward propagation neural networks with image preprocessing [18] and scale invariant feature transform [19] and convolution deep neural network [20].

The purpose of the work is to carry out a comparative analysis of the two methods of traffic signs recognition and identify their application features, as well as to formulate possible modifications of the given methods to achieve the best quality of recognition.

2. Materials and methods
This study utilizes a modular approach to creating a computer vision system with an option of separate optimization and adjustment. The system considered is a collection of two modules, the tasks of which are distributed as follows: module I – selecting areas that may contain a sign; module II – detecting and classifying signs in the selected area.

Schematic representation of the sequence of actions (interaction of system modules) is shown in figure 1.

To use the first approach, a three-layer perceptron with a system selecting ROI on the basis of color filtering. The system of selecting the ROI consists of the following stages:
1) filtering noise using a Gaussian filter;
2) converting the image from the RGB color space to the HSV space to facilitate searching under different illumination conditions;
3) applying color filters to select areas containing basic colors of the signs in question;
4) applying morphological operations for noise filtering of the color filter (erosion) and restoring the shape of the sign with filling areas of non-basic colors (dilatation);
5) removing areas not matching in size and aspect ratio.

Areas found through this procedure are scaled to 30x30 pixels in size. For each such area, a descriptive vector consisting of 63 values is calculated. For its calculation, each of the found areas is converted into a grayscale image, then the average brightness value of all its pixels is calculated. First 30 components of the vector consist of the number of pixels in each image row, brightness of which is above the average value. The following 30 components consist of the number of pixels in each image column, brightness of which is above the average value. The remaining 3 components of the vector represent average values of color components in the HSV color space in the found area.

Classification of the detected areas is carried out with the help of a three-layer perceptron having 63 neurons in the first layer, 80 neurons in the hidden layer and the number of neurons in the output layer being equal to the number of signs considered. A sigmoid is used as an activation function in the network. A normalized describing vector of each region is sent to the input of the network.

In the second approach, in order to determine the ROI a convolution neural network with a small number of filters is used to increase performance. Neural network structure is shown in figure 2.
Figure 2. Convolution neural network structure for module I.

As the network input, the image from the DVR with high resolution, converted to a size of 1200x600, is used. The output of the network is an array of data of size 200x100 with a range of possible values from 0 to 1, what represents a probability scale of finding a sign in the region corresponding to the area in the original image.

As a loss log, the intersection over union function was used (1). The main optimization method was the Adam stochastic optimization approach [21].

\[
IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}
\]

After receiving the probability array, the ROI’s are selected from the original image (to preserve the high resolution) and transferred to the input of the module II.

To implement module II in this approach, a fully connected convolutional neural network with indirect communication between convolution layers was developed. The network structure is shown in figure 3.
Figure 3. Convolution neural network structure for module II.
Due to the need to perform the tasks of detecting and classifying an image, the learning selection and network structure were presented in the form of multiple naming. Such a representation allows to annotate an image using an output vector elements of which are disconnected and can take a high and a low values independently of other output elements. To implement this representation, a list of signs being classified was drawn up: a stop sign, a bus stop sign, a pedestrian crossing sign and a main road sign.

Thus, the network output vector consists of four neurons. The output activation function was chosen as sigmoidal, which is a smooth nonlinear function with limiting values from 0 to 1, what corresponds to the probability of classifying the inclusion of the sign with the image.

As an estimation function, a binary crossentropy function (2) was chosen.

$$F(t, p) = -\sum_{i} (t_i \log(p_i) + (1 - t_i)\log(1 - p_i))$$

Such a function is often used in logistic regression problems. Adagrad was chosen as an optimization approach [22].

3. Results

In order to compare the considered approaches, we use the basic metrics applied for the evaluation of classification problems: recall (3), precision (4), F1 (5).

$$R = \frac{TP}{TP + FN},$$

$$P = \frac{TP}{TP + FP},$$

$$F1 = \frac{2 * R * P}{R + P},$$

TP being the number of correct classifications of the presence of a sign on the frame, FN being erroneous classification of sign's absence on the frame, FP being erroneous classification of sign's presence on a frame.

When implementing the first approach for learning a neural network, a training sample was created from 255 domains containing the whole sign or part of it, as well as areas that do not contain signs, but have a similar color. Neural network learning was carried out using the backpropagation algorithm. The results of network's test using a test selection are presented in Table 1.

| Metrics       | Sign          |
|---------------|---------------|
|               | Stop          | Bus stop       | Pedestrian crossing | Main road |
| Recall        | 0.8           | 0.333          | 0.625               | 1.0       |
| Precision     | 1.0           | 0.833          | 0.882               | 0.846     |
| F1-score      | 0.888         | 0.476          | 0.732               | 0.917     |

In the second approach, a test set of 500 annotated frames from a dash cam was created for a convolution neural network of module I to learning. 20% of the images were randomly picked from the selection for network validation. Also, to reduce the probability of network's overtraining, data augmentation of training set was used by applying basic image processing operations: move, rotate, expand, compress.

As a result of learning, the value of estimation function reached 0.63, while the value of network recall was 0.86, what actually was the main purpose of module I learning, i.e. predicting which frame areas contain a sign. However, precision is not evaluated.

An example of image annotation using a convolution neural network is shown in figure 4.
Figure 4. Example of frame annotation using a neural network.

For the convolution neural network to learn and get configured in module II, a selection of 400 image sections with signs names depicted on frames was selected and annotated. Also, a selection of 40 images was created for testing. Neural network learning for module II was also carried out using artificial data extension method. The selection for validation was 20% of the training set.

According to the results of learning, error function reached 0.001 in the learning selection and 0.01 in the validation selection. The results of network's test using a test selection are presented in table 2.

Metric values were chosen on the basis of the largest F1 value of the metric when the threshold for sign labeling changed. The threshold of sign labeling means the smallest value of activation of the output neuron, at which the labeling of the sign value of this neuron is considered active. The value of 0.6 was chosen as the threshold for sign labeling.

| Metrics | Sign         |
|---------|--------------|
|         | Stop         | Bus stop     | Pedestrian crossing | Main road |
| Recall  | 0.8          | 0.866        | 1.0                | 0.8       |
| Precision | 0.88888    | 1.0         | 0.909              | 0.888     |
| F1-score | 0.84        | 0.92        | 0.95               | 0.84      |

4. Conclusion
As a result of comparing two approaches for implementing the module of signs classification, the following features were revealed:

- the approach using image preprocessing is sensitive to changes in environmental conditions;
- both approaches have a high robustness in when classifying signs with noises;
- the approach using a convolution neural network requires a wider selection to enhance the ability to correctly annotate values without overtraining, what in its turn in the approach with color filtering falls on adding and modifying algorithms;
- the approach utilizing color filtering has a higher potential for scaling and implementation compared to convolution networks due to a lower requirements for computational resources;
- extension in convolution networks becomes available through increasing the number of filter layers.
- the following possible modifications are planned in future:
- module I to utilize properties of recurrent networks in view of the work with a continuous video stream;
• using a sliding frame inside the section containing a sign for localizing the sign;
• prioritizing signs for the driver (depending on the position on the image).

References
[1] Shneier M 2005 Proc. IEEE Computer Society Int. Conf. on Computer Vision and Pattern
Recognition pp 215 – 222
[2] Ruta A, Porikli F, Li Y, Watanabe S, Kage H and Sumi K 2009 IAPR Conference on Machine
Vision Applications (MVA) Session 15: Machine Vision for Transportation
[3] Belaroussi R, Foucher P, Tarel J P, Soheilian B, Charbonnier P and Paparoditis N 2010 20th
International Conference on Pattern Recognition (ICPR) pp 484 – 488
[4] Houben S, Stallkamp J, Salmen J, Schlipsing M and Igel C 2013 International Joint Conference
on Neural Networks
[5] Stallkamp J, Schlipsing M, Salmen J and Igel C 2012 Neural Networks 32 pp 323 – 332
[6] Erhan D, Szegedy C, Toshev A and Anguelov D 2014 CVPR
[7] Fergus R and Zisserman P A 2007 IJCV 71 pp 273 -303
[8] Bouchard G and Triggs B 2005 CVPR, IEEE Computer Society pp 710 -715
[9] Felzenszwalb P, McAllester D and Ramanan D 2008 CVPR
[10] Bosch A, Zisserman A and Munoz X 2007 CVIR
[11] Chum O and Zisserman A 2007 CVPR
[12] Dalal N and Triggs B 2005 CVPR pp 886-893
[13] Ferrari V, Fevrier L., Jurie F and Schmid C 2008 PAMI 30 pp 36-51
[14] Rowley H A, Baluja S and Kanade T 1996 NIPS 8 pp 875-881
[15] Viola P and Jones M 2001 CVPR 1 p 511
[16] Lampert C H, Blaschko M B and Hofmann T 2008 CVPR
[17] Li L, Wu Y and Ye M 2014 Informatica (2013) 38 pp 145–153
[18] Auranuch L and Jackrit S 2017 International Journal of Innovative Research in Computer and
Communication Engineering 5
[19] Hu X, Zhu X, Lia D and Li H 2010 A special joint symposium of ISPRS Technical Commission
IV & AutoCarto in conjunction with ASPRS/CaGIS 2010 Fall Specialty
ConferenceNovember 15-19
[20] Ciresan D, Meier U, Masci J and Schmidhuber J 2012 Multi-Column Deep Neural Network for
Traffic Sign Classification.
[21] Kingma D and Ba J 2015 3rd International Conference for Learning Representations
[22] Duchi J, Hazan E and Singer Y 2011 Journal of Machine Learning Research 12 pp 2121-2159