Robust Chinese Traffic Sign Detection and Recognition with Deep Convolutional Neural Network

Rongqiang Qian, Bailing Zhang, Yong Yue

Department of Computer Science
Xi'an Jiaotong-Liverpool University
Suzhou, China

Zhao Wang

Department of Computer Science
Bournemouth University
Bournemouth, UK

Frans Coenen

Department of Computer Science
The University of Liverpool
Liverpool, UK

Abstract—**Detection and recognition of traffic sign, including various road signs and text, play an important role in autonomous driving, mapping/navigation and traffic safety. In this paper, we proposed a traffic sign detection and recognition system by applying deep convolutional neural network (CNN), which demonstrates high performance with regard to detection rate and recognition accuracy. Compared with other published methods which are usually limited to a predefined set of traffic signs, our proposed system is more comprehensive as our target includes traffic signs, digits, English letters and Chinese characters. The system is based on a multi-task CNN trained to acquire effective features for the localization and classification of different traffic signs and texts. In addition to the public benchmarking datasets, the proposed approach has also been successfully evaluated on a field-captured Chinese traffic sign dataset, with performance confirming its robustness and suitability to real-world applications.**

Keywords-component; traffic sign detection; traffic sign recognition; deep learning;convolutional neural networks; multi task CNN

I. INTRODUCTION

Acquisition of the information from various traffic signs is crucial in many applications, such as autonomous driving, mapping and navigation. It is also important in intelligent transportation systems. Generally, a traffic sign recognition system involves two related issues: traffic sign detection and traffic sign classification. The former aims to accurately localize the traffic signs in an image, while the later intends to identify the labels of detected object into specific categories/subcategories. Though the topic has attracted research interests in computer vision community for many years [1], it is generally regarded as challenging due to various complexities, for example, diversified backgrounds of traffic sign images. On the other hand, Chinese traffic signs are much more complex compared with western countries due to the large number of Chinese characters. How to efficiently detect and recognize the traffic signs in China has virtually not been discussed.

With appropriate approaches of object detection for traffic signs from images, the mainstream traffic sign recognition is a two-stage procedure, namely, feature extraction followed by classification. Many feature description algorithms have been proposed for traffic sign recognition, for example, circle detector [2], histogram of gradients, scale-invariant feature transform (SIFT) and Haar-wavelet [3]. However, these manually engineered features are separated with the classifier design, which means sub-optimal system performance as there is no joint optimization for the two modules.

Recently, the development of deep learning has attracted much attention in computer vision research as more and more promising results are published on a range of different vision tasks. Among the deep learning models, the convolutional neural networks (CNN) have acquired unique noteworthiness from their repeatedly confirmed superiorities. The CNN models have also been applied to traffic sign recognition, for instance, committee CNN [4], multi scale CNN [5], multi column CNN [6] and hinge-loss CNN [7].

The main characteristics of our proposed traffic sign recognition system include a multi-task CNN model to learn a compact yet discriminative feature representation, which simultaneously implement detection and classification, and a color-based region proposal for the improvement of the famous R-CNN model [8].

The rest of this paper is organized as follows. In section 2, some recent research on traffic sign detection, traffic sign recognition and CNN applications will be introduced. Then, the detail of our proposed multi-task CNN model will be presented in section 3, 4 and 5. Finally, experiments and some discussions are given in section 6 and 7, respectively.

II. RELATED WORK

Many works have been published in recent years for traffic signs detection and recognition. The classical approaches generally treated traffic sign detection and recognition differently, which is in sharp contrast with current deep learning methodology.

A. Traffic sign detection

Traffic sign detection is similar to other object detection tasks in computer vision, namely, identifying the image regions with bounding boxes that tightly contain a traffic

sign. Compared with other objects in a traffic image, traffic signs are usually associated with salient shapes, and uniform and distinctive colors. The positional relationship with other visual objects along the road is also an important cue that could be exploited in the development of traffic sign detection. Based on how the information is used, the literature on traffic sign detection can be roughly categorized into geometry-based, segmentation-based or hybrid [9].

There are two schools of thought in object detection, namely, sliding window paradigm exemplified by the AdaBoost proposed by Viola and Jones [10] and HOG+SVM [11], and object proposal that aims to produce a set of regions (i.e., object proposals) that have high probability to contain objects [8]. Following the lines, there are corresponding approaches proposed for traffic sign detection. The majority of published works are in the former category. Bahlmann et al. [12] proposed a real-time detection scheme based on AdaBoost. Inspired by generalized Hough transform, the orientation and intensity of image gradient are used for traffic sign detection [13]. The application of canny edge detection in Ruta et al.'s work [14] reported a perfect detection rate. Houben et al. [15] proposed a Hough-like algorithm for detecting circular and triangular shapes. Timofte et al. [16] use integer linear programming to learn a set of good colour transformations together from the training data with the optimal threshold.

Aiming to overcome the inherent limitations of sliding window approach, the object proposal has gained much attention recently. Greenhalgh employed Maximally Stable Extremal Regions (MSERs) for traffic sign detection [17]. Salti proposed a pipeline which is based on the extraction of interest regions and experimentally confirm the advantages over some sliding window approaches [9].

B. Traffic sign recognition

Traffic sign recognition can be generally treated as a pattern classification issue, with many mature off-the-shelf techniques from machine learning. Among the plenteous models, support vector machine (SVM) is the most popular one, which has been applied in [18, 19]. Boosting is another popular method for traffic sign classification. Ruta et al. [3] proposed a robust sign similarity measurement with SimBoost and fuzzy regression tree for road sign recognition. Baró et al. [20] trained an ensemble of classifiers in the Error-Correcting Output Code (ECOC) framework, where the ECOC was designed through a forest of optimal tree structures that are embedded in the ECOC matrix.

As for any visual object classification, feature expression is the critical factor for the system performance. How to design discriminative and representative features has been in the central stage of computer vision research. In the last decade, a plethora of image description algorithms has been proposed, many of which have also been attempted for traffic sign recognition. For example, the HOG and Haar-wavelets were used in Stallkamp et al's work [21], scale invariant feature transform (SIFT) were applied in [22, 23]. Liu et al. [24] presented a feature learning approach using group sparse coding and achieved a good result on GTSRB dataset [21]. However, it requires a good deal of

time to design such features and it is difficult to promote the robustness to specific task.

C. Convolutional neural network

Deep learning [25,26] generally refers to representational learning hierarchical features from data, which is in marked contrast with hand crafting features as SIFT or HOG. Recent advancement in deep learning and deep Convolutional Neural Networks (CNN) in particular, has produced a wide range of outstanding results on object detection and recognition benchmarks. The most important attributing factor for the success is the end-to-end framework, which integrates feature extraction and classification or other computer vision tasks.

Krizhevsky et al. [27] proposed a large-scale deep convolutional network trained by standard back propagation, with breakthrough on the large-scale ImageNet object recognition dataset [28], attaining a significant gap compared with existing approaches that use shallow models. Girshick et al. [8] used selective search [29] to propose candidates region, and trained a high-capacity CNN model, in which supervised pre-training and domain-specific fine-tuning are applied, achieving 30% mean average precision (mAP) improvement on PASCAL VOC dataset than previous best result. Wang et al. [30] built an end-to-end text recognition system in natural scene by two CNNs, one for detection, the other for classification, achieving state-of-the-art performance.

With the availability of some large benchmark traffic sign datasets, such as BTSRB [16] and GTSRB [21] datasets, there have been a number of empirical studies for traffic sign detection and classification with the aid of deep learning. Several CNN-based methods [4-7] have been proposed, reporting state-of-the-art performance with the two public datasets. Ciresan et al. [31] presented a multi-column approach for image classification, which takes average of the outputs from several deep neural networks (or columns) trained on inputs preprocessed in different ways. Later the model was applied to traffic sign classification [6], achieving recognition rate 99.46% on the GTSRB dataset. A hinge loss stochastic gradient descent method was proposed [7] to train CNNs, and further improved the record on the GTSRB to 99.65%. In [32], the so-called extreme learning machine (ELM) classifier was applied together with CNN for traffic sign recognition, which leverages the discriminative capability of deep convolutional features and the outstanding generalization performance of ELM classifier, obtaining competitive results on the GTSRB dataset.

III. SYSTEM OVERVIEW

Generally, traffic signs are designed with regular shapes and attractive color to be easily noticed. The most widely used traffic signs can be classified into three subcategories, namely, prohibitory type, including circle, red rim, white or red inner; mandatory type, including circle, blue rim, blue inner; and danger type, including triangular, black rim, yellow inner. Some of the samples are illustrated in Figure 1. Traffic sign recognition is more challenge in China since the traffic signs are not only limited to the three common types

of traffic signs, but also include digits, English letters and Chinese characters.

Figure 1. Samples of Chinese traffic signs

Our work on traffic sign detection and recognition was mainly inspired by one of the most notable CNN models, namely, "regions with convolutional neural network" (R-CNN), proposed by Girshick et al. [8] for object detection, which has demonstrated state-of-the-art performance on standard detection bench-marks. By selective search method, a few hundreds or thousands candidate bounding boxes will be produced as the proposal regions. This may not be always efficient or even necessary. For traffic sign detection, the uniform and distinctive colors can be exploited to simplify the generation of proposal regions.

There are mainly two modules in our proposed system, as demonstrated in Figure 2. The first module generates region proposals based on a much simplified yet efficient scheme for producing candidate regions. The main characteristics of our approach are by RGB Space Thresholding, followed by Connected Component Analysis (CCA). The second module is the CNN model to discriminate and classify each of the candidate regions. For each regions produced, a multi-task CNN will be trained to determine if a candidate region proposal is traffic sign or not, and reject false samples accordingly. At the same time, the specific categories of the true samples will also be determined.
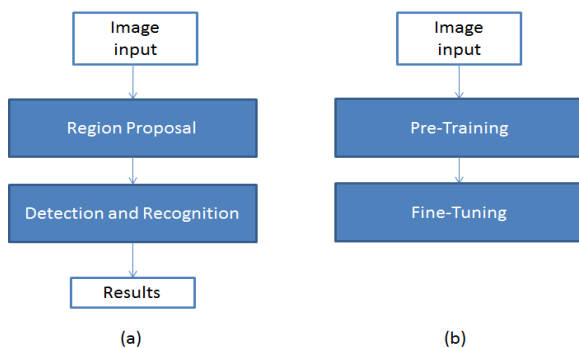

Figure 2. System overview. Fig. 2(a) displays the testing procedures. Fig. 2(b) shows the training steps.

In next section, we will present more details for each module, describe their test-time usage, and explain how their parameters are learned.

IV. REGION PROPOSAL

In our system, an input RGB image is first converted into a set of binary images by applying a set of pre-set thresholds for each of the R, G and B channels. The motivation for generating multiple binary images from multiple channels is to compensate for the large illumination variability. From our experience, a range from 10 to 240 with an interval of 10 is sufficient for threshold adaptation. In other words, the minimum and maximum values of the threshold are 10 and 240 respectively, within which the threshold will be recursively updated for the binarization. Thereafter, edge detection and Connected Component Analysis (CCA) will be performed on each binary image to produce the candidate regions. Figure 3 provides an illustrative example.
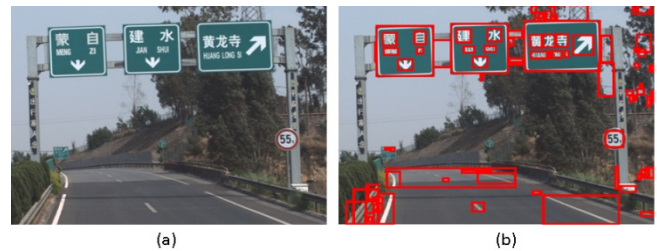

Figure 3. Region Proposal. Fig. 3(a) displays the original image. Fig. 3(b) illistruates the bounding boxes of all the proposed regions

V. CONVOLUTIONAL NEURAL NETWORKS

We applied the CNN model with the similar structure as in [4], which can be illustrated in Figure 4. The network consists of three convolution stages followed by fully connected layers and two final softmax layers. Each convolution stage consists of convolutional layer, non-linear activation layer and max pooling layer. All the pooling layers used are max pooling. ReLU [27] is employed as the activation function for convolutional layers and full connection layers. The final pooling layer is shared by two fully connected layers, i.e., multiple layer perceptron (MLP), corresponding to prediction and classification, respectively.
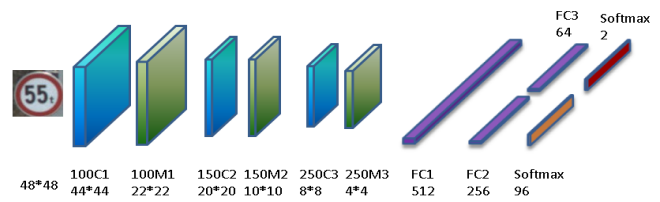

Figure 4. CNN architecture

The first MLP is designed for detection, with 64 hidden units and 2 output units for positive/negative decision. The second MLP is designed for the classification of specific categories of a detected object. There are 96 classes of object, including traffic signs, digits, English letters and Chinese

characters with class number of 40, 10, 26, and 20 respectively.

The structure of the networks and the hyper-parameters were empirically initialized based on previous works using ConvNets [4], then we setup cross-validation experiment to optimize the selection of network architecture. The detail CNN parameters are shown in Table I.

TABLE I. CNN PAPAMETERS

Layer	Type	Feature maps & Size		kernal
1	Input	1 map with 48×48 neurons		
2	Convolution	100 maps with 44×44 neurons		5×5
3	Max pooling	100 maps with 22×22 neurons		2×2
4	Convolution	150 maps with 20×20 neurons		3×3
5	Max pooling	150 maps with 10×10 neurons		2×2
6	Convolution	250 maps with 8×8 neurons		3×3
7	Max pooling	250 maps with 4×4 neurons		2×2
8	Fully connection	512 neurons		
		Task 1	Task 2	
9	Fully connection	64 neurons	256 neurons	
10	Softmax	2 neurons	96 neurons	

VI. EXPERIMENT

A. Performance evaluation

To evaluate the performance of our system, a computer with CPU Xeon 3.3GHz and 16GB memory is employed. For the purpose of increasing training speed, a very efficient Titan GPU based implementation was performed by using NVIDIA CUDA and CUDNN. Six different datasets were employed, which are described together with the corresponding experiments in the following.

B. Pre-training datasets: GTSRB, MNIST and CASIA GB1

Three dataset are employed for pre-training of the proposed CNN.

GTSRB [21]: The German Traffic Sign Recognition Benchmark has 43 classes of traffic signs. The training data set contains 39209 training images in 43 classes and the test data set contains 12630 test images.

MNIST [33]: The MNIST database of handwritten digits has a training set of 60,000 examples come from 10 classes, and a test set of 10,000 examples.

CASIA GB1 [34]: Institute of Automation of Chinese Academy of Sciences (CASIA in GB1 set), which contains 300 samples for each of 3755 characters.

We reconstructed a 96 class dataset with total number 283304 by collecting 43 classes from GTSRB, 10 classes from MNIST and 43 classes from CASIA GB1. The detail pre-training scheme is as follows:

1. Initial weights of convolution layers and full connection layers are achieved from a uniform random distribution in the range [−0.05, 0.05].

2. Pre-train the deep model by using the 96 classes from previous dataset. The network is pre-trained as a classifier with single task. We adopt the cross-entropy loss and stochastic gradient descent (SGD) in back propagation.

3. The learning rate is 0.001 with weight decay of 1% for each epoch. 100 epochs are trained in pre-training stage.

C. Fine-tune datasets: Field-captured Dataset

The fine-tune dataset was recorded by a camera set up in vehicle under different time and shooting angle. The captured videos have a size of 1600×1236 pixels. We then extract 96 classes with total number 100000 samples (Figure 5) from the captured videos. The distributions between positive and negative samples are 60% and 40%, as shown in Table II.

Figure 5. Samples of Field-captured dataset

TABLE II. FIELD-CAPTURED DATASET

	Positive	Negative	Number
Training dataset	48000	32000	80000
Testing dataset	12000	8000	20000
Total	60000	40000	100000

All the weights and biases have been pre-trained in the previous stage. Aiming to perform detection and recognition tasks simultaneously, multi-task CNN is used. Therefore, we add one full connection layer and one softmax layer. More specifically, the additional layers will execute the function of detection, as Figure 4 illustrates.

When multi-task back propagation is performing, losses are minimized by following linear combination:

$$L = \sum_{i=1}^{2} w_i L_i \qquad (1)$$

where w_i are weights for each task. In our implementation, weights of both the tasks are set to 1.

In this stage, the learning rate is also 0.001 with weight decay of 1% for each epoch. Cross-entropy loss and stochastic gradient descent (SGD) are used for executing back propagation, and the network usually coverages within 60 epochs.

Once fine-tune is finished. We use testing dataset to examine the performance. The correct rate for detection and

recognition are 97.56% and 98.83%, respectively, which is indeed encouraging. The result also indicates that the performance does not influence much for different situations.

D. LP Dataset [35]

LP dataset is a detection dataset, it contains 410 Chinese vehicle images (Figure 6) and bears varied imaging conditions such as resolution, illumination and viewing angles. Digits, English letters and Chinese characters are involved in this dataset.

Figure 6. Samples of LP dataset

The metric [35] of evaluation for the two benchmark datasets was defined by: (i) high level-true, i.e., license plate is totally encompassed by the bounding box and $A \cap B / A \cup B \geqslant 0.5$, where A is the detected region and B is the ground truth region; (ii) low level-false, i.e., the license plate is totally missed by the bounding box; and (iii), middle level-partial, namely the remaining results excluded by the above two types.

TABLE III. LP DATASET

Approach	Accuracy		
	True	Partial	False
Proposed	90.2%	7.4%	2.4%
PVW [35]	93.2%	0.3%	6.5%
LPE [35]	84.6%	1.5%	13.9%
HLPE [35]	80.8%	6.6%	12.6%
ESM [35]	74.6%	7.3%	18.1%

As Table III illustrates, it is obvious that our algorithm achieves the lowest false detection rate. Moreover, a competitive true detection rate is also obtained.

E. GTSDB

The GTSDB dataset is combined by a training dataset and a test dataset, which contains 600 and 300 images respectively. Three categories of traffic signs are included in these images. Specifically, they are Prohibitory, Mandatory and Danger. Although the sematic definitions of traffic signs are different from each other, traffic signs belong to one target category has some common properties. In instance, Prohibitory signs have circular red borders, Danger signs have triangular red borders and Mandatory signs have blue backgrounds and white arrows, as shown in Figure 7.

Figure 7. Samples of GTSDB

By applying the proposed method, we achieved a competitive performance of 95% detection accuracy for GTSRB.

VII. CONCLUSTION

In this paper, a multi-task CNN based road traffic information acquisition method is proposed. Aiming to detect and recognize not only traffic signs, but also digits, English letters and Chinese characters. The whole procedure includes two stages. For any input image, firstly, a set of candidate regions are proposed by using colour space thresholding. Secondly, multi-task CNN is used to determine the similarity and reject false samples in detection task. Simultaneously, the detail categories of the true samples are obtained by classification task. The approach is evaluated on several popular datasets, achieving comparative results.

Further extension of this work is suggested to focus on: 1) enhancing the performance of region proposal; and 2) classification by using part-based.

REFERENCES

[1] H. Akatsuka, S. Imai, Road signposts recognition system, Tech. rep., SAE

[2] M. Meuter, C. Nunny, S. M. Gormer, S. Muller-Schneiders, A. Kummert, Adecision fusion and reasoning module for a traffic sign recognition system, Intelligent Transportation Systems, IEEE Transactions on 12 (4) (2011) 1126–1134.

[3] A. Ruta, Y. Li, X. Liu, Robust class similarity measure for traffic sign recognition, Intelligent Transportation Systems, IEEE Transactions on 11 (4) (2010) 846–855.

[4] D. Ciresan, U. Meier, J. Masci, J. Schmidhuber, A committee of neural networks for traffic sign classification, in: Neural Networks (IJCNN), The 2011 International Joint Conference on, IEEE, 2011, pp. 1918–1921.

[5] P. Sermanet, Y. LeCun, Traffic sign recognition with multi-scale convolutional networks, in: Neural Networks (IJCNN), The 2011 International Joint Conference on, IEEE, 2011, pp. 2809–2813.

[6] Dan Cireşan et al, Multi-column deep neural network for traffic sign classification, Neural Networks, Vol. 32, August 2012, Pages 333–338.

[7] J. Jin, K. Fu, C. Zhang, Traffic sign recognition with hinge loss trained convolutional neural networks, Intelligent Transportation Systems, IEEE Transactions on 15 (5) (2014) 1991–2000

[8] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, 2014, pp. 580–587.

[9] S. Salti, A. Petrelli, F. Tombari, N. Fioraio, L. D. Stefano, Traffic sign detection via interest region extraction, Pattern Recognition, Volume 48, Issue 4, April 2015, Pages 1039-1049, ISSN 0031-3203.

[10] P. Viola, M. Jones, Robust real-time object detection, International Journal of Computer Vision 4 (2001) 34–47.

[11] G. Wang, G. Ren, Z. Wu, Y. Zhao, and L. Jiang, "A robust, coarse-to-fine traffic sign detection method," in Proceedings of IEEE International Joint Conference on Neural Networks, 2013.

[12] C. Bahlmann, Y. Zhu, V. Ramesh, M. Pellkofer, T. Koehler, A system for traffic sign detection, tracking, and recognition using color, shape, and motion information, in: Intelligent Vehicles Symposium, 2005. Proceedings. IEEE, IEEE, 2005, pp. 255–260.

[13] G. Loy, N. Barnes, Fast shape-based road sign detection for a driver assistance system, in: Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on, Vol. 1, IEEE, 2004, pp. 70–75.

[14] A. Ruta, Y. Li, X. Liu, Real-time traffic sign recognition from video by classspecific discriminative features, Pattern Recognition 43 (1) (2010) 416–430.

[15] S. Houben, A single target voting scheme for traffic sign detection, in: Intelligent Vehicles Symposium (IV), 2011 IEEE, IEEE, 2011, pp. 124–129.

[16] R. Timofte, K. Zimmermann, L. Van Gool, Multi-view traffic sign detection, recognition, and 3d localisation, Machine vision and applications 25 (3) (2014) 633–647.

[17] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide-baseline stereo frommaximally stable extremal regions, in: Image Vision Comput. 22 (2004) 761–767.

[18] S. Maldonado-Bascón, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gómez-Moreno, F. López-Ferreras, Road-sign detection and recognition based on support vector machines, Intelligent Transportation Systems, IEEE Transactions on 8 (2) (2007) 264–278.

[19] M. Shi, H. Wu, H. Fleyeh, Support vector machines for traffic signs recognition, in: Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, IEEE, 2008, pp. 3820–3827.

[20] X. Baró, S. Escalera, J. Vitrià, O. Pujol, P. Radeva, Traffic sign recognition using evolutionary adaboost detection and forest-ecoc

classification, Intelligent Transportation Systems, IEEE Transactions on 10 (1) (2009) 113–126.

[21] J. Stallkamp, M. Schlipsing, J. Salmen, C. Igel, Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition, Neural networks 32 (2012) 323–332.

[22] F. Zaklouta, B. Stanciulescu, Real-time traffic-sign recognition using treeclassifiers, Intelligent Transportation Systems, IEEE Transactions on 13 (4) (2012) 1507–1514.

[23] J. Greenhalgh, M. Mirmehdi, Real-time detection and recognition of road traffic signs, Intelligent Transportation Systems, IEEE Transactions on 13 (4) (2012) 1498–1506.

[24] H. Liu, Y. Liu, F. Sun, Traffic sign recognition using group sparse coding, Information Sciences 266 (2014) 75–89.

[25] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks., Science (5786) (2006) 504.

[26] G. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Computation 18 (7) (2006) 1527–1554.

[27] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei- Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 248–255.

[29] J. Uijlings, K. van de Sande, T. Gevers, A. Smeulders, Selective search for object recognition, International Journal of Computer Vision 104 (2) (2013) 154–171.

[30] T. Wang, D. Wu, A. Coates, A. Ng, End-to-end text recognition with convolutional neural networks, in: Pattern Recognition (ICPR), 2012 21st International Conference on, 2012, pp. 3304–3308.

[31] D. Cires an, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In CVPR, 2012.

[32] Yujun Zeng, Xin Xu, Yuqiang Fang, Kun Zhao, Traffic Sign Recognition Using Extreme Learning Classifier with Deep Convolutional Features, The 2015 International Conference on Intelligence Science and Big Data Engineering (IScIDE 2015), Suzhou, June 14-16, 2015.

[33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, November 1998.

[34] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang. Chinese Handwriting Recognition Contest. In Chinese Conference on Pattern Recognition, 2010.

[35] W. Zhou, H. Li, Y. Lu, , Q. Tian, Principal visual word discovery for automatic license plate detection, IEEE Transactions on Image Processing 21 (9) (2012) pp. 4269–4279.