Recognizing Text-based Traffic Guide Panels with Cascaded Localization Network

Xuejian Rong[1], Chucai Yi[2], and Yingli Tian[1]

[1]The City College, City University of New York, New York, NY 10031, USA
[2]HERE North America LLC, Chicago, IL 60606, USA

{xrong, ytian}@ccny.cuny.edu, gschucai@gmail.com

Abstract. In this paper, we introduce a new top-down framework for automatic localization and recognition of text-based traffic guide panels[1] captured by car-mounted cameras from natural scene images. The proposed framework involves two contributions. First, a novel Cascaded Localization Network (CLN) joining two customized convolutional nets is proposed to detect the guide panels and the scene text on them in a coarse-to-fine manner. In this network, the popular character-wise text saliency detection is replaced with string-wise text region detection, which avoids numerous bottom-up processing steps such as character clustering and segmentation. Text information contained within detected text regions is then interpreted by a deep recurrent model without character segmentation required. Second, a temporal fusion of text region proposals across consecutive frames is introduced to significantly reduce the redundant computation in neighboring frames. A new challenging Traffic Guide Panel dataset is collected to train and evaluate the proposed framework, instead of the unsuited symbol-based traffic sign datasets. Experimental results demonstrate that our proposed framework outperforms multiple recently published text spotting frameworks in real highway scenarios.

Keywords: Road Scene Understanding, Traffic Guide Panel, Text Spotting, Video OCR

1 Introduction

With the recent advances in vehicle intelligence, advanced driver assistance, and road surveying, the vehicle mounted systems are expected to have a deep understanding of the surrounding environment and provide reliable information for the drivers or autonomous navigation. As one of the most important context indicators in driving status, traffic signs (symbol-based or text-based) have attracted considerable attention in the fields of detection and recognition. Symbol-based traffic signs such as *Stop* or *Exit* signs usually have relatively smaller size and

[1] http://tinyurl.com/wiki-guide-signs

unique shape, while text-based traffic signs/panels often have a standard rectangular shape containing numerous text information.

Most existing algorithms for traffic sign recognition were developed specifically for symbol-based traffic signs under limited conditions. Some systems demanded sophisticated hardware setup to capture high-resolution images, and others worked well on individual static frames but could not meet the efficiency requirement of the real-time video processing. Moreover, these methods usually ignored a large amount of valuable semantic information resided in the text-based traffic signs such as guide signs or panels, as shown in Figure 1. The semantic information from guide panel could notify drivers or autonomous control systems of interchange, toll plaza, and exit direction. In most cases, this information is not completely available or up-to-date on car-mounted navigation systems.

Fig. 1. Samples of traffic guide panels and the corresponding text information in the highway environments.

In this paper, a framework is proposed to detect and recognize text-based traffic guide panels captured in highway environments (see examples in Figure 1). This framework could help deliver the text information from guide panels to human drivers as head-up display information, and also to autonomous driving vehicles in case of un-updated digital mapping. On a set of continuous image frames, we first detect candidate traffic guide panels in each frame by using a set of learned convolutional neural network (CNN) features, and then eliminate false positive candidates by using temporal information from the continuous frames.

The preliminary guide panel candidates are then further enhanced. Afterward, a fine CNN-based text detector is trained to localize all the detected text regions within the guide panel candidates. These text regions are finally recognized by a deep recurrent model in a sequence-to-sequence encoder-decoder fashion. Although several general scene text detection and recognition methods [1,2,3] have been developed to localize and recognize the text information in the natural scenes, most of these methods used exhaustive manner such as sliding window to search for all possible regions containing text information across an entire image. This process is time-consuming and error-prone, leading to more false alarms. The shape, color, and geometric cues of the traffic signs are not completely modeled in these approaches. In contrast, our proposed framework could largely reduce the searching space in each frame and improve the efficiency.

The remainder of the paper is organized as follows: Section 2 reviews related work on traffic sign reading and text spotting in the wild. Section 3 describes our methodology for localizing the traffic guide panels and corresponding scene texts, and the enhancement of the guide panels. The temporal information fusion of consecutive video frames, and the recognition of the detected text regions are described in details in Section 4. The collected highway traffic guide panels benchmark dataset and the experimental results are presented and discussed in Section 5. The proposed framework and future work are concluded in Section 6. It is worth noting that the single-word based guide signs (e.g., *STOP* sign) are out of the scope of this paper since their inside text information is always fixed and they could be directly detected and recognized as specific types of traffic signs.

2 Related Work

In recent years, many researchers worked on the research topics associated with text extraction from natural scene images and its associated applications. Most state-of-the-art methods of scene text extraction [1,2,4,5,6,7,8,9,10] comprise two stages, detection to obtain image regions containing text information, and recognition to transform image-based text information into text codes. The detection methods could be further divided into three groups: region-based methods, e.g., [10], connected component based methods, e.g., [11,12,13], and convolutional neural network (CNN) based methods, e.g., [8]. Region-based text detection approaches rely on local features like the texture to locate text regions, while connected component-based methods focus on segmenting individual text characters using specific text patterns such as the intensity, colors, and edges. And CNN-based approaches usually attempt to generate the character saliency maps based on the extracted CNN features on multiple scales, and apply clustering afterward.

For the text recognition, a number of techniques [9,14,15] have been reported which follow a bottom-up fashion to train their own scene character classifiers. The recognized characters are then grouped into a word based on the context information, while some errors including spelling and ambiguities are recovered by

incorporating the Lexicon and n-gram language model. Most of these methods require robust and accurate character-level segmentation and recognition. To avoid the above numerous local computation, several methods based on recurrent neural network (RNN) with long short-term memory (LSTM) are recently proposed [16,17], which model a word image as an unsegmented sequence and does not require character-level segmentation and recognition.

Although scene text extraction has been a fairly popular research field, there are a limited number of publications that specifically concentrated on the extraction of text information from traffic guide signs and panels captured in the form of continuous frames by car-mounted cameras. The main challenges which prevent the exploration might be the wide diversity of the information contained within traffic panels which are difficult to analyze, and the computation complexity of the popular text extraction approaches which cannot meet the efficiency requirement in realistic environments.

Specifically, Gonzlez et al. [18] attempted to use maximally stable extremal regions (MSERs) to detect both traffic signs and text characters. The traffic panels were detected in each frame based on color segmentation and bag of visual words, and the detected regions were further classified using both support vector machines and Naive Bayes classifiers. However, this method was only applied to the single frame and ignored the temporal information. Greenhalgh et al. [19] introduced more scene cues like the scene structure to define search regions within each frame, and exhaustedly located a large number of guide sign candidates using MSERs, hue, saturation and value color thresholding. The text characters contained within detected candidate regions are further detected as MSERs and grouped into lines, before being interpreted with optical character recognition (OCR) engines. This approach outperforms previous methods, but is still sophisticated and computationally expensive. Notice, these methods mainly only focus on locating the general traffic signs and rely on the OCR engines for the following text detection and recognition inside the guide signs. However, our proposed method attempts to provide more accurate and useful text region proposals and their interpretation besides the basic guide panel locations which are relatively meaningless to the drivers or autonomous driving systems.

For the system validation, several datasets have also been proposed, including the German traffic sign detection benchmark [20], the German traffic sign recognition benchmark [21], and the Belgian traffic sign dataset [22]. However, these datasets focus on the detection of the symbol-based traffic sign, and therefore not applicable to the validation for extracting text information from the guide panel.

3 Cascaded Localization of Text-based Guide Panels

In this section, the localization process of the traffic guide panels and the text regions within are presented. The enhancement of the preliminarily localized guide panels is also described. Specifically, to accurately and efficiently localize the guide panels and text regions of interest, we establish a two-stage cascaded

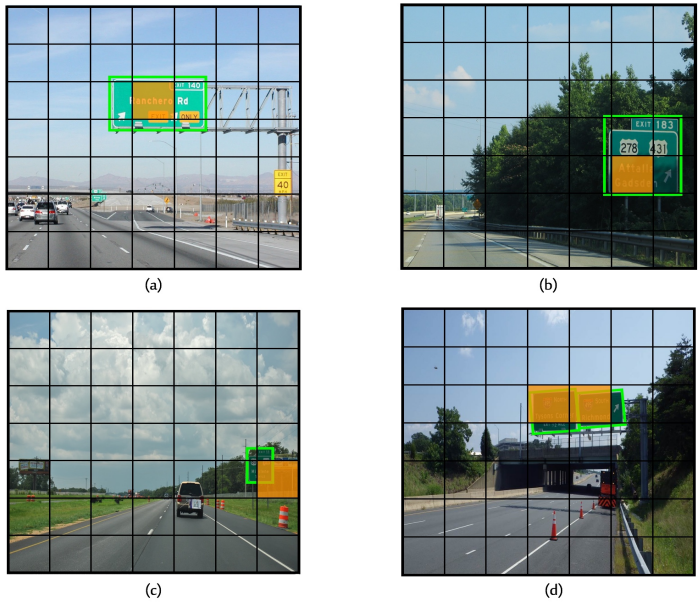


Fig. 2. Demonstration of localizing guide panel candidates. The highlighted grids in orange represent where the center of the guide panel falls into, and predict the exact shape of the corresponding bounding box. The green bounding box represents the final regressed shape of the panel prediction. The input resolution of the first layer of the guide panel localization net is 448×448. And the total number of grids is $7 \times 7 = 49$.

framework, and model each stage as a unified detection process, inspired by the You Only Look Once (YOLO) detector [23]. The first stage of the proposed Cascaded Localization Network (CLN) aims to find all the guide panels candidates with a high recall rate, and the second stage focuses on the accurate localization of the text regions and eliminates the false alarms, including the non-panel and redundant detections, with the text localization results.

3.1 Unified Localization of Traffic Guide Panels

For an image frame captured at the highway, we first search for all the possible locations of traffic guide panels by integrating the separate components of object detection into a single neural network. Therefore the network reasons globally about the full image and all the candidates, and predicts all the bounding boxes simultaneously. In practice, the detector first evenly divides the input highway scene image (rescaled to 448×448) into a $S \times S$ grid, in which each grid cell is responsible for detecting and localizing the guide panel whose center falls into this grid cell, by predicting B bounding boxes and confidence scores for those boxes. The confidence scores represent the probability of the box containing a guide panel and also the accuracy of the box prediction. The confidence is formally defined as $Pr(Panel) * IoU^{GT}_{Pre}$, which would be forced to zero if there

is no object existing in the cell, and otherwise equal the intersection over union (IoU) between the predicted box and the ground truth.

Different from the original YOLO detector, here each bounding box is composed of 7 predictions: $\{x, y, w, h, \cos\theta, \sin\theta\}$ and the presence confidence. The (x, y) coordinates and the width/height (w, h) denote respectively the location (center of the box w.r.t. the bounds of the grid cell) and size of a bounding box tightly enclosing the guide panel. θ represents the bounding box rotation. And the presence confidence denotes the IoU between the predicted guide panel bounding box and any ground truth bounding box. The implementation details of this localization network are as follows. We set $S = 7, B = 2$ in the experiments on the newly collected Traffic Guide Panel dataset. The final prediction is a $S \times S \times (B * 7 + 1) = 7 \times 7 \times 15$ tensor. To boost the efficiency of the localization process, this network has 9 convolutional layers followed by 2 fully connected layers w.r.t. the 24 convolutional layers used in regular YOLO detector. We pretrain our convolutional layers on the ImageNet 1000-class competition dataset [24], and fine-tune the model on the training set of the Traffic Guide Panel dataset including the ground truth annotations for all the text-based traffic guide panels. The localization results are illustrated in Figure 2.

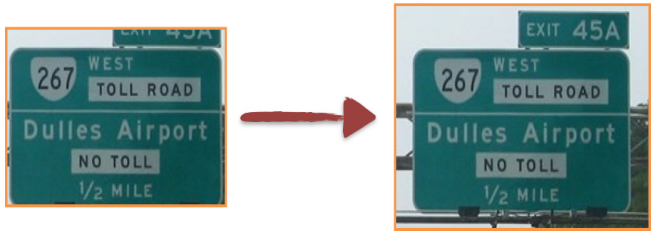


Fig. 3. Illustration of the enhancement of initial localized guide panels by extending each edge of the bounding box with 20%.

Compared with previously CNN-based object detection approaches which attempt to generate the saliency map and verify the clustered bounding box for specific object category, here the regression-based guide panel detector directly outputs the localization results without extra panel/non-panel classification. The generated bounding boxes of guide panels are then enlarged by 20% to each side, as illustrated in Figure 3, since the predicted panel bounding boxes, are sometimes too tight which will affect the following text region localization. In addition, this enlargement is able to make up missing parts of guide panels from small bounding boxes. This process also involves context information of the guide panels to benefit the following recognition stage.

3.2 Fine Text Region Localization

In this section, we introduce the second stage CNN architecture of the cascaded localization network for text region localization in the enhanced guide panel

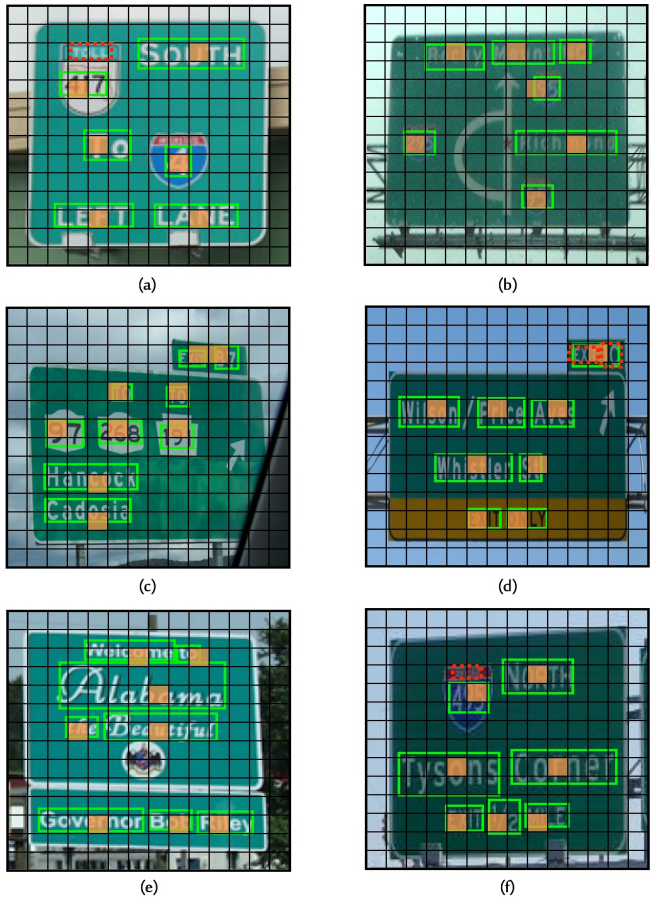


Fig. 4. Localization results of the text region candidates within detected traffic guide panels. The highlighted grids in orange predict the exact shape of the bounding box, and represent where the center of the text region candidates fall into. The green bounding boxes represent the finally regressed shape of the text region prediction. The red dashed bounding boxes represent the missed ground truth text region.

image patches. This stage also follows the strategy of the YOLO detector and its variant [25] by constructing a fixed field of predictors, each of which specializes in predicting the presence of a word string around a specific image location. To better localize the text regions instead of the general objects, additional Hough Voting predictor is implemented to pool evidence from the whole image. Since the text occurrences are usually smaller and more variable compared to general objects, the grid number is also increased from $G_I = 7$ to $G_I = 14$ to solve the problems. In details, the architecture comprised 9 convolutional layers, each of which is followed by the Rectified Linear Unit non-linearity (ReLU), four of

which by a max-pooling layer (2×2, with a stride of 2). The stride of all linear filters is 1, and the resolution of feature maps is kept through zero padding.

In testing each grid cell is responsible to localize a word string if the word center falls within this cell, by regressing five numbers: the text region presence confidence c, and the other four parameters $p = (x, y, w, h)$, where the (x, y) coordinates represent the text location (center of the box w.r.t. the bounds of the grid cell), and the width and height (w, h) represent the size of the text region. Here we do not need to predict the bounding box rotations of multiple text regions as in [25], since all the text lines are parallel to the edge of the guide panels based on observation. Therefore, for an input guide panel patch of size $H_I \times W_I$, we obtain a grid of $(H_I \times W_I)/\Sigma^2$ predictions, one each for an image cell of size $\Sigma \times \Sigma$ pixels. To effectively detect the large text instances in the guide panel, we further apply the text region localization net on multiple scales. The input image is first downsampled by factors $\{1, 1/2, 1/4\}$, and the localization results at multiple down-scaled levels of the input image are finally merged via non-maximal suppression. In two overlapping detections, the one with the lower probability is suppressed. The final localization results are demonstrated in Figure 4.

4 Recognizing Extracted Text Regions

On the localized guide panels and their text regions from a sequence of continuous frames, text recognition is performed to extract readable text codes. The temporal information across continuous frames is modeled to reduce the computation cost and eliminate the false alarms. This fusion step is unnecessary if the proposed localization pipeline is applied on a static image or individual frame. A deep recurrent model is then introduced to directly recognize all the localized text regions.

4.1 Temporal Fusion in Consecutive Frames

In the practical driving process, the proposed cascaded localization networks should run on each of the continuous frames captured from the car-mounted camera/recorder, and the occurrence of the guide panels would be relatively rare in a long frame sequence. Moreover, the resolution and quality of the guide panel would gradually increase as car proceeded close and then suddenly vanish. Therefore, it would be computationally time-consuming and unnecessary to apply all the cascaded localization stages for every video frame as many previous traffic sign detection methods [13,18,19].

In our experiments, only the guide panel localizer is applied on each frame, and the enhancement and text region localizer are only applied in the last t frames, where the actual value of t is determined by the trade-off between the effective text recognition time and average text recognition accuracy. The last available frame is determined by the distance between the guide panel image

patch and the three image boundaries (top, left, and right). Some specific localization results are regarded as false alarms if they are not successfully localized by the guide panels at the last t frames. In our test, $t = 10$ works well for the fusion process, which usually takes $1/6 \sim 1/3$ of the whole time of successful panel detection in the highway environments.

Fig. 5. The sequence-to-sequence encoder-decoder recognition demonstration of the extracted text region candidates.

4.2 Recognizing Text Regions with Deep Recurrent Model

After the finally refined text regions are generated, the text recognition process is modeled as a sequence labeling problem by using a deep recurrent model. In the traditional framework of text recognition in traffic signs or license plates, character segmentation has a great influence on the success of recognition. The text information would be recognized incorrectly if the segmentation is improper, even if a high-performance character recognizer is adopted. Here we model the text region recognition problem as a single-pass sequence labeling process. In details, each input text region x ideally contain a piece of text with horizontal orientation from left to right. The overall procedure of the sequence labeling based text region recognition starts with converting the text region bounding box into a sequence of feature vectors which are extracted by using the pre-trained 9-layer CNN model sliding across the bounding box. Afterward, a bi-directional recurrent neural network (RNN) model with long short-term memory (LSTM) is trained to label the sequence features, with stochastic gradient descent (SGD) algorithm. Connectionist temporal classification is applied at last to the output layer of RNN to analyze the labeling results of RNN and generate the final recognition result. The recognition process is demonstrated in Figure 5.

5 Experimental Results

In this section, experiments are performed to verify the effectiveness of the proposed Cascaded Localization Network on new Traffic Guide Panel dataset which contains a variety of individual highway guide panels, compared with multiple recent text spotting approaches in the realistic highway scenes.

Benchmark Dataset Since there is no publicly available dataset specifically on traffic guide panels, we collect a new challenging dataset of traffic guide panels at the highway. This dataset contains a variety of highway guide panels {3841 high-resolution individual images in total, 2315 containing traffic guide panel level annotations (1911 for training and 404 for testing, and all the testing images are manually labeled with ground truth tight text region bounding boxes), 1526 containing no traffic signs}. All the images are collected from *AAroads* website[2], and captured from the view of car-mounted dash camera, including numerous kinds of traffic guide panel such as direction, toll plaza, destination distance, and exit indication.

Fig. 6. Comparison of the Top-5 text region localization proposals from the proposed approach and the best competing baseline method [8].

[2] `http://www.aaroads.com`

In the experiments, all the traffic guide panel annotations in 1911 of the 2315 images are used to fine-tune the guide panel localization net. The text region localization net is trained on the SynthText in the Wild Dataset [25], which consisted of 800k images with approximately 8 million synthetic word instances. Each text instance is annotated with its text-string-level, word-level, and character-level bounding boxes.

Table 1. Text localization results and average processing times on the Traffic Guide Panel dataset. Precision P and recall R at the maximum f-measure F, and the localization time t_l (in seconds).

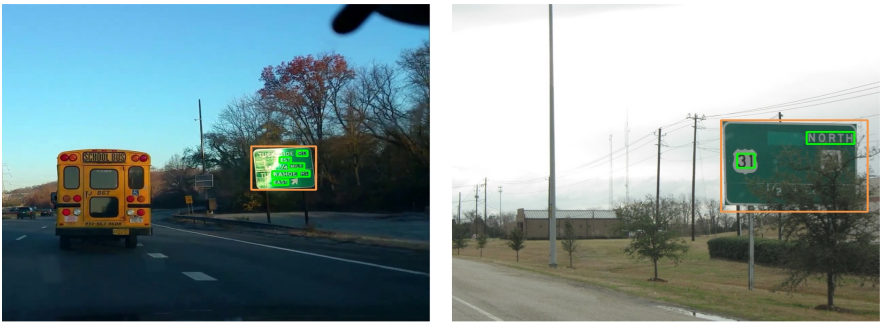
Method	P	R	F	t_l
Proposed	**0.73**	0.64	**0.68**	0.16
Jaderberg et al. [8]	0.59	**0.71**	0.64	4.53
Gomez et al. [26]	0.46	0.53	0.49	1.32
Epshtein et al. [11]	0.35	0.41	0.38	2.51

Comparison to existing methods First our proposed approach is compared with three recent methods for lexicon-free text detection.

- Stroke Width Transform, Epshtein et al. [11]: a well-known method[3] that leverages the consistency of characters' stroke width to detect arbitrary fonts.
- MSER Text Detection, Gomez et al. [26]: uses maximally stable extremal regions (MSERs), a popular tool in text detection[4], which is combined with a perceptual organization framework.
- Deep Text Spotting, Jaderberg et al. [8]: a state-of-the-art method[5] that uses multiple stages of convolutional neural networks to predict text saliency score at each pixel, and cluster to form the region predictions afterward.

For the first two methods, the outputs are ranked by the bounding box size, which is a sensible way to favor the more prominent detected texts since the codes do not produce confidence values. For [8], the summed text saliency scores are used for candidates ranking. Table 1 shows the text localization performance and computation efficiency (i.e., average processing time) on a standard PC with dual 3.2 GHz CPU and a NVIDIA Geforce Titan X GPU. We follow the standard PASCAL VOC detect criterion: a detection is correct if the IoU between its bounding box and the ground truth exceeds 50%. Overall, our method outperforms the existing text localization methods in the highway environments, and the gains over the two non-learning methods [11,26] are large

[3] `https://github.com/lluisgomez/DetectText`
[4] `https://github.com/lluisgomez/text_extraction`
[5] `https://bitbucket.org/jaderberg/eccv2014_textspotting`



(a) (b)

Fig. 7. Failure cases of the proposed Cascaded Localization Networks due to kinds of image degradations, e.g., reflection and occlusion.

in terms of f-measure. Moreover, the proposed method outperforms the conventional R-CNN based text detection approach [8] on the precision and f-measure, and is comparable in terms of recall rate. As to the computation efficiency, due to the straightforward and precise regression architect, the proposed cascaded localization network performs significantly faster than the previous learning and non-learning methods.

Qualitative examples Finally, we present text detection examples in Figure 6 to qualitatively demonstrate the performance of the proposed approach and the best competing baseline [8]. These images illustrate the advantages of our proposed method for narrowing down the search space and improving the computation efficiency. Failure cases in certain frames caused by image degradations, such as uneven illumination, reflection, and occlusion, are demonstrated in Figure 7. However, these localization results could be effectively eliminated through temporal fusion in practice.

6 Conclusion and Future Work

In this paper, we have presented a new top-down CNN-based cascaded framework for automatic detection and recognition of text-based traffic guide panels in the wild. The proposed framework performed in an efficient coarse-to-fine manner, and effectively reduced the redundant computation in continuous frames. The future work will focus on further improving the accuracy and efficiency of the cascaded localization network on traffic guide panels, and extending the newly collected text-based guide panel dataset to a larger scale for future validation and comparison.

Acknowledgement This work was supported in part by NSF grants EFRI-1137172, IIP-1343402, and FHWA grant DTFH61-12-H-00002.

References

1. Ye, Q., Doermann, D.: Text Detection and Recognition in Imagery: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(7) (2015) 1480–1500
2. Yin, X., Zuo, Z., Tian, S., Liu, C.: Text Detection, Tracking and Recognition in Video: A Comprehensive Survey. IEEE Transactions on Image Processing **25**(6) (2016) 2752–2773
3. Zhu, Y., Yao, C., Bai, X.: Scene text detection and recognition: recent advances and future trends. Frontiers of Computer Science **10**(1) (2016) 19–36
4. Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., Bai, X.: Multi-Oriented Text Detection with Fully Convolutional Networks. arXiv.org (April 2016)
5. Qin, S., Manduchi, R.: A Fast and Robust Text Spotter. WACV (2016)
6. Zini, L., Odone, F.: Portable and fast text detection. Machine Vision and Applications (2016) 1–15
7. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading Text in the Wild with Convolutional Neural Networks. International Journal of Computer Vision **116**(1) (2015) 1–20
8. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep Features for Text Spotting. ECCV **8692**(Chapter 34) (2014) 512–528
9. Neumann, L., Matas, J.: Efficient Scene Text Localization and Recognition with Local Character Refinement. arXiv.org (2015)
10. Yin, X., Yin, X., Huang, K., Hao, H.: Robust Text Detection in Natural Scene Images. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(5) (2014) 970–983
11. Epshtein, B., Eyal, O., Yonatan, W.: Detecting text in natural scenes with stroke width transform. CVPR (2010)
12. Yi, C., Tian, Y., Arditi, A.: Portable Camera-Based Assistive Text and Product Label Reading From Hand-Held Objects for Blind Persons. IEEE/ASME Transactions on Mechatronics **19**(3) (2014) 808–817
13. Wu, W., Chen, X., Yang, J.: "Detection of text on road signs from video". IEEE Transactions on Intelligent Transportation Systems (Dec 2005)
14. Wang, T., Wu, J., Coates, A., Ng, A.: End-to-end text recognition with convolutional neural networks. ICPR (2012)
15. Rong, X., Yi, C., Yang, X., Tian, Y.: Scene text recognition in multiple frames based on text tracking. (2014)
16. Su, B., Lu, S.: Accurate Scene Text Recognition Based on Recurrent Neural Network. ACCV (2014)
17. Shi, B., Wang, X., Lv, P., Yao, C., Bai, X.: Robust Scene Text Recognition with Automatic Rectification. arXiv.org (March 2016)
18. Gonzalez, A., Bergasa, L., Yebes, J.: "Text detection and reocnition on traffic panels from street-level imagery using visual apperance". IEEE Transactions on Intelligent Transportation Systems **15**(1) (Feb 2014) 228–238
19. Greenhalgh, J., Mirmehdi, M.: "Real-time detection and recognition of road traffic signs". IEEE Transactions on Intelligent Transportation Systems **16**(3) (Dec 2012) 1360–1369
20. Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: "Detection of traffic signs in real-world images: The German traffic sign detection benchmark". in Proc. IJCNN (Aug 2013) 1–8

21. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: "The German Traffic Sign Recognition Benchmark: A multi-class classification competition". in Proc. IJCNN (July 2013) 1453–1460
22. Timofte, R., Zimmermann, K., Gool, L.V.: "Multi-view traffic sign detection, recognition, 3D localisation". WACV (2009)
23. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. CVPR (2016)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: "Imagenet classification with deep convolutional neural networks". NIPS (2012) 1097–1105
25. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic Data for Text Localisation in Natural Images. arXiv.org (April 2016)
26. Gomez, L., Karatzas, D.: Multi-script text extraction from natural scenes. ICDAR (2013)