Deep Active Learning Approach for Traffic Sign and Panel Guide Arabic-Latin Text Content Annotation in Natural Scene Images

SUBMISSION DATE / POSTED DATE

30-05-2022 / 02-06-2022

CITATION

Zaghdoud, Ridha; Boukthir, khalil; Hamdani, Tarek M.; Alimi, Adel M. (2022): Deep Active Learning Approach for Traffic Sign and Panel Guide Arabic-Latin Text Content Annotation in Natural Scene Images. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.19929383.v1

DOI

10.36227/techrxiv.19929383.v1

Deep Active Learning Approach for Traffic Sign and Panel Guide Arabic-Latin Text Content Annotation in Natural Scene Images

Ridha Zaghdoud[a,b], Khalil Boukthir[b], Tarek M. Hamdani[b], Adel M. Alimi[b,c]

[a] *University of Sousse, ISITCom, 4011 Sousse, Tunisia*
[b] *REsearch Groups in Intelligent Machines (REGIM Lab), University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia.*
[c] *Department of Electrical and Electronic Engineering Science, Faculty of Engineering and the Built Environment, University of Johannesburg, Johannesburg, South Africa*

ABSTRACT

The detection and recognition of road traffic signs and panel guides content has become challenging in recent years. Few studies have been made to solve these two issues at the same time especially in Arabic language. Additionally, the limited number of datasets for traffic signs and panel guide content makes the investigation more interesting. In our work, we propose a Deep Active Learning Approach for Traffic Sign and Panel Guide Arabic-Latin Text Content Annotation in Natural Scene Images. Convolution neural network (CNN) and active learning are combined in the proposed system to detect and recognize traffic sign and multilingual scene text from panels guides, particularly those with Arabic Latin characters. The annotation system based on active learning method is applied to the Natural Scene Traffic Sign and Panel Guide Arabic-Latin Text dataset (NaSTSArLaT) of Tunisian highway road. The defined dataset contains initially 3000 collected images with only 181 annotated samples. Progressively, un-annotated training images, are automatically annotated when they provide high confidence with the YOLOv5 detection model and only the less confident examples are manually annotated. The annotation output consists of a class label and its bounding box for every traffic sign and panel guide text content. The performance of the proposed system is done on test data that represents 10% of the available annotated set. The proposed active learning strategy gives 62.7% in terms of mean Average Precision (mAP) only with the manual annotation of about ¼ of the samples, whereas if the learning is performed using all the data after a complete annotation, the obtained performance is 69.2% which is comparable to our approach.

Keywords: Active learning, Deep learning, Annotation, Traffic sign Dataset, Natural scene images, Text detection

1. Introduction

During the previous century, Self-driving automobile research has considerably progressed. [1]. The development of artificial intelligence (AI) has been accelerating advancements in automobile intelligence and technology for driverless cars, opening a novel era of transit wherein automobiles are capable to perceive and understand the environment using varying degrees of automation to achieve every task. In this new generation, the drivers will continue to play an essential role as well as cars, will become the intelligent partners that offer individualized guidance. Within this context, speech controllers allow the conductor to monitor certain functions and settings of the car in place of the traditional way of commanding by hand, the best voice control technology like Apple Siri, Amazon Alexa [2], Microsoft Cortana, Bixby and Google Assistant are examples for this conventional technique. Their navigation systems interact with central datasets to determine where traffic congestion exists and which direction to drive to avoid them.

e-mail: ridha.zaghdoud@regim.usf.tn (Ridha Zaghdoud), khalil.boukthir@regim.usf.tn (Khalil Boukthir), tarek.hamdani@regim.usf.tn (Tarek M. Hamdani), adel.alimi@regim.usf.tn (Adel M. Alimi)

The car may also be able to survey the situation of the driver and the traffic (facial expressions, eye gestures, voices, traffic signs, gates, etc.) to launch a warning device or take over the conduct when drivers are feeling tired or stressed. These progresses in the automatic sector are principally made possible by the various strategies to include artificial intelligence (AI) into the work production process and the way in which vehicles interact with their surroundings.

As an important part of automated driving, traffic prediction plays a crucial role in traffic state management. Road sign identification in scene images has lately been a matter of study. Despite of the large volumes of available algorithms, the majority of them concentrates on traffic signs and excludes other signs with textual information (e.g. panel guide). On practice, however, the many types of markers, varied types of data, and advanced textual dialects make text reading in traffic signs a difficult process. Furthermore, the availability for difficult naturalist datasets is limited with a variety of signs and writing dialects hamper the progress of further gradual techniques of detecting textual signs and regulatory.

The manual image annotation technique still more accurate, but with the increase in the quantity of image data the manual method stays inefficient in terms of cost, time and human annotators With the integration of automatic labelling technology, it can significantly reduce costs, and with some degree of human intervention, accuracy is guaranteed. This research presented a semi-automatic annotation approach based on a data preparation system that can achieve good integrity and accuracy while minimising human works.
In the process of creating and annotating datasets,Calling only the manual method is insufficient to annotate a very large number of complicated traffic scenes, which would be far from meeting the requirements of large-scale data study and analysis, and is also an impossible task.

Three contributions are presented in this work:

(1) In contrast with many specific datasets in the literature, the current work presents a benchmark that contains text in natural scene images that corresponds specifically to the road vocabulary including Arabic writings in highway panel content guide.

(2) Due to the problem of annotation time with large available amount of unlabeled data, we use semi-automatic annotation to specify which instances to label in order to achieve high accuracy model calling as few as possible human assistance.

(3) To test state of the art text detection and object detection deep learning techniques, we propose a multi-purposes dataset that covers Traffic Sign and Panel Guide Arabic-Latin Text Content.

The remainder of the paper is split into three sections. After Introduction We'll start with a description on the different dataset that exist, both at the text level and at the traffic sign level and some works on how annotation of datasets is done. In section 3, we are going to describe the sequence to follow to reach this dataset starting with the collection until the annotation at two levels manual and semi automatic with active learning. In section 4, experiment and results are described in details. Finally, the conclusion section is a summary of the work provided.

2. Traffic Sign, Panel Guide Text Content and Semi-Annotated Dataset: Related works

In this section we will present some datasets in which we will indicate those that use manual annotation and those that use semi-automatic annotation.

2.1. Traffic Sign Datasets

In order to aid in the advancement of reliable road sign recognition methods, many benchmark datasets have been suggested. Among the biggest and most various datasets for road

signs comprehension we can mention The **German Traffic Sign Recognition Benchmark** (GTSRB) [7], This benchmark was gathered on German highways and includes over 50000 pictures of signs in different shapes and sizes (between 20 and 250 pixels), organized into 43 categories.

German Traffic Sign Detection Benchmark (GTSDB) [11], It is a benchmark consisting of 900 images devised in two parts, for the train, 600 images were taken, and for the test there are 300 images, all images are taken in the roads of Germany.

LISA [8], is a video and captioned image collection of American road signs. It includes 7855 images of traffic signs with 47 different types.

Tsinghua-Tencent 100k [10], the famous dataset for detecting street signs in China. This is widely recognized as one of the most powerful data collections ever assembled, delivering 100,000 images including 30000 examples of traffic signs, with an image resolution of 2048 to 2048.

These types of datasets were created solely for the purpose of detecting and recognizing traffic signs and are hence unsuitable for extracting text from road signs, including guidance signs. To tackle the problem of not having traffic sign Datasets available, **Rong et al** [12] gathered a collection of difficult data on highway traffic signs in the United States.

Gonzalez et al [9], Google Street View was used to build a dataset of 2 distinct Spanish motorways that they applied to validate their retrieval approach.

ASAYAR [6], a new multilingual dataset published in 2020 focused to the detection of Latin (French) and Arabic scene text on roadway signs. It contains around 1763 photos with detailed annotations. Moroccan Road provided the dataset, which have been manually labeled and using 16 classes.

To summarize, table 1 contain a detailed comparison between the different existing datasets.

2.2. Panel Guide Text Scene Detection Datasets

A variety of text datasets have been used to help the work on text scene recognition.

TSVD [5], this collection contains 7K photos of Tunisian cities with Arabic text images taken from Google Street View. The active learning technique reduces picture volumes for annotation, which is a benefit of this dataset. Although this dataset

Table 1: Comparison of Existing Datasets.

Detection Dataset	ICDAR17[3]	SVT[4]	TSVD[5]	ASAYAR[6]	GTSRB[7]	LISA[8]	Gonzales et al[9]	Tsighua-Tencent100k[10]	NaSTSArLaT
Traffic Sign		×		×	×	×	×	×	×
Traffic Sign with Arabic Word Text	×	×		×					×
Traffic Sign with Latin word Text	×	×		×	×	×			×
Deep Active Learning based Annotation			×						×

contains a wide range of Arabic scene text, only around a fifth of the training examples have been tagged and used.

ARASTI [13], Texts from Arabic scenes in images, Arabic scene segmented words, and Arabic scene segmented characters may be found in this Dataset. It contains pictures of diverse naturalistic environments that have been separated between 2093 segmented Arabic characters and 1280 segmented Arabic sentences cropped from 374 scene images.

MSRA-TD500 [14], it is made up of 200 testing pictures and 500 training images with randomly curved English and Chinese texts.

ICDAR2017-MLT [3], ICDAR2017- MLT is a vast, multilingual text collecting project that includes 7200 training examples, 1800 validation pictures, and 9000 images for testing. The data collection contains images of a natural scene with text in nine languages.

Street View Text [4], is a collection of 725 annotated word pictures obtained from Google Street View (SVT) which gathered information from 350 pictures.

2.3. Semi-annotated Datasets

There are several available datasets that have been manually annotated(see table 1). As a result, text in images scene is annotated into separate text lines [[15], [16], [13]], whereas we quote[[6], [9]] for annotating objects.

On the other side The automation of data collecting and labeling is described in some of the works studied in this research in order to improve accuracy and minimize the time spent on data annotation.

The technique of Ke et al. [17] is based on machine learning frameworks that create textual descriptions of images using a pretrained CNN (what classes of objects are in the image). This method outperformed prior automatic image labeling with text techniques and outperformed manual data annotation.

Song et al. [18] mention automatic image annotation as well. In order to extract data from images more correctly, they used an intermediary layer in a neural network. The method provides for precise textual descriptions of images, but it is not adaptable: as new data emerges, the model must be retrained on a new dataset, which must be manually annotated. We also quote AcTiv Dataset in which Oussama et al. [19] present a semi-automatic technique.

3. Deep Active Learning based Annotation for NaSTSArLaT Dataset

In this section we present the steps to follow to obtain in the end a dataset consisting of several categories of images and suitable to test it with several models of text detection or objects detection. We also present a statistic about this dataset by seeing in figure 2.

3.1. Dataset Challenges

The variety and complexity of panel guide or scene texts is a difficult mission on processing this dataset, these issues can be mostly developed in three aspects:

- The variety and changeability of texts in natural scenes: Unlike scripts in documents, the natural scene text has a much greater diversity and variability. For example, A scene's text can be in a variety of languages, colors, font sizes, rotations, and forms.
- Background complexity and interference: The backgrounds of natural scenes are almost uncontrollable. Trees are also obstacles in the scene, buildings, vehicles or even signs that obscure other signs.
- Imperfect imaging environment: The text instances can be found corrupted as well as the taken image often suffering from motion blur and this decreases text recognition performance.

3.2. Data acquisition process

The images are obtained on Tunisian highways from Google Street View developed by Google where witch are offer two versions of the images captured from the year 2016 and 2021 with different resolutions, our own have a resolution of 6656*3328. We have chosen the distance between two positions 300m.The number of images downloaded is 50000 panoramic images, some of them do not contain signs. The retrieved images include a variety of Tunisian highway signs, comprising in particular; signs with text-based information in Arabic/French that we have a large percentage of the images, Priority road signs, Directional, Prohibitory, Danger and Temporary (see in figure 1). The dimension of the images was adjusted to 1000×500, with 96 dpi transverse and longitudinal resolution.

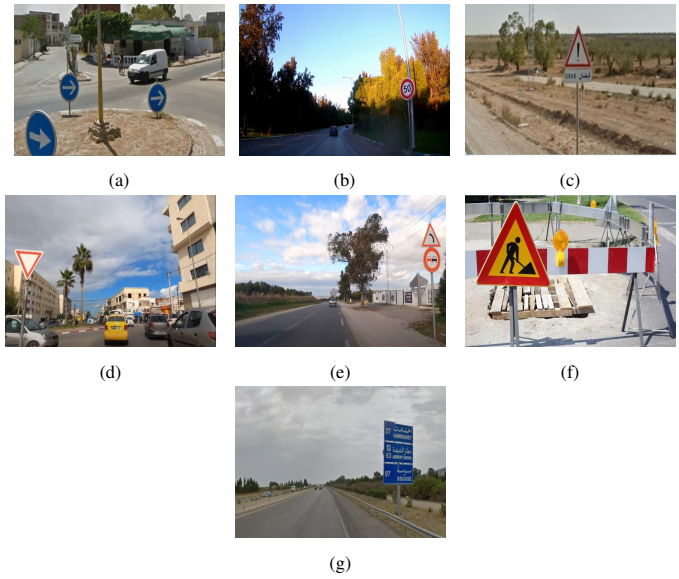


Fig. 1: Examples of images containing road signs from our Dataset. (a)Directionally sign (b) Regularity sign(c) Danger sign (d) Priority sign (e) Prohibitory (f) Temporary (g) Panel Guide

3.3. Data Annotation process

We have two steps for the annotation first a part of is manually annotated by Roboflow, this tool automatically generates a package of textual information in YOLOv5 PyTorch, is an interesting tool to use it helps developers to structure classify and create datasets. You can also share the work between a workgroup and communicate between them online and integrate, directly into your modeling process. We split our dataset into two categories:

- Data of Traffic Sign: contains the different types of road signs
- Data of Textual Traffic: contains panel guide on Arabic and Latin language

As mentioned in table 2, NaSTSArLaT contains 3000 images with a total of number of 13065 samples, on average 4.4 boxes per image, which we find the big part by Data of textual traffic with 7903 samples which proves the major existence of the panel guide, the total are distributed as shown on figure 2.

3.3.1. Data Traffic Sign

The driving in Tunisia is well adopted thanks to the variations of the road signs. We find a traffic road sign that gives recommendations and made aware of any risk or danger may be encounter the driver in road (see in figure 1). We have 10 classes for annotating traffic road as we can see in figure 2.

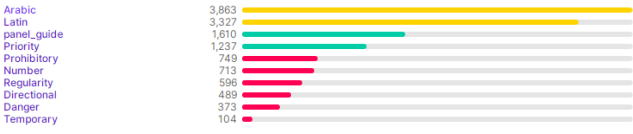


Fig. 2: Classes balance

3.3.2. Traffic Text

These types of signs show the direction of the countries and their names in French and Arabic, as well as the number of kilometers to the destination mentioned.

We have focused mainly on the text existing in the signs, this type of traffic is a means of communication between users and the road. Data of traffic contains the following categories: Arabic, Latin and Number. Some examples are mentioned in figure 4.

3.3.3. Deep active learning annotation

The main idea comes from [20], which suggested a suggestive annotation technique. Because of the large identification demand for varied object, it's a challenge to minimize annotation costs without compromising detection efficiency. With a small completely labeled dataset and a big unlabeled dataset, offers a weakly supervised learning strategy for object detection. Both labeling expenses and storage space may be saved using our weakly labeling strategy.

In the first place, the user will annotate a small dataset (set 1) which will then be divided into three sub-sets by bootstrapping techniques which randomly generates training data each of which will be trained by an object detection model, which in our case is the YOLOv5, after which the most efficient model will be selected, The next step is to test our model on a set of un-annotated images, the images which have confidence above 60 percent are not affected and then check the ones that have no objects detected, which in turn will be annotated and added to the initial dataset to reset the process from the beginning. (see in figure 3)

4. EXPERIMENTS RESULTS ON NaSTSArLaT Dataset

In this part, we show the usefulness of our NaSTSArLaT Dataset for object detection and text detection in panel guide, we also show the optimum model's performance. To divide our base we estimate it in two parts starting with training (90 %) and validation (10 %). Then we will see the effect of our approach of active learning to facilitate the task of annotation in order to have a good dataset with small amount of data and reduce the time needed to accomplish this task.

4.1. Evaluation metrics

In computer vision mAP is a very popular evaluation measure used in detection whether it is in the localization of instances known by bounding boxes or classification, several object detection algorithms like Faster-RCNN and YOLO use mAP to evaluate their performance.

Average Precision (AP): is a metric that evaluates accuracy by combining recall and precision. AP calculation is defined as

$$AP = \sum_{i=1}^{n-1} (r_{i+1} - r_i)\, p_{iterp}\,(r_{i+1}) \tag{1}$$

Mean Average Precision (mAP): the most frequently used measure in research papers AP is calculated for each class and

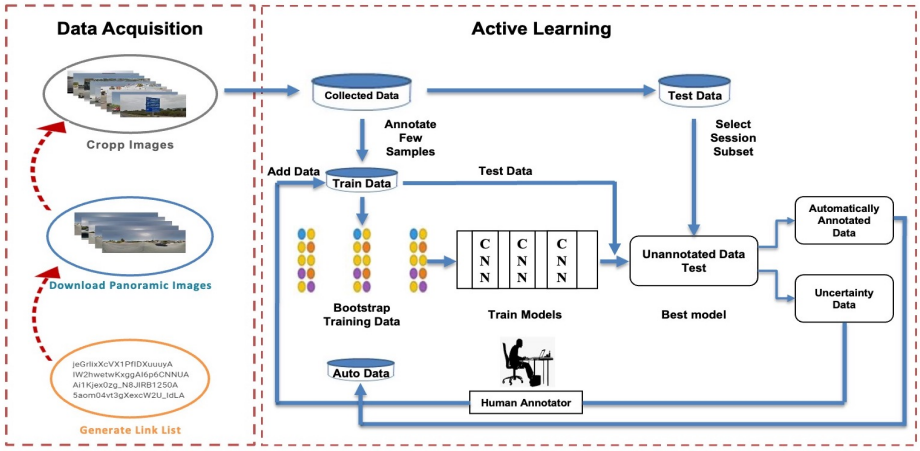
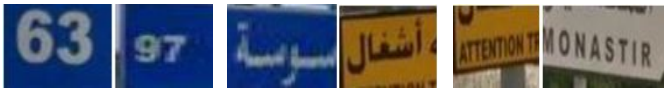


Fig. 3: Framework annotation based on deep active learning

Table 2: NaSTSArLaT Dataset distribution

Categories	Classes	Number of Samples
Data of Traffic Sign	7	5158
Data of Textual Traffic	3	7903



(a) (b) (c)

Fig. 4: Examples of textual traffic categories (a) Number (b) Arabic (c) Latin

(a) (b) (c) (d) (e) (f)

Fig. 5: Visual representation of the test results on traffic sign using deep learning approach.

average all result for all object to obtain the mAP, which will encapsulate all predictions into a single value. The formula for calculating it is shown in equation.

$$mAP = sum_{k=1}^{k=n} AP_k/n \quad (2)$$

Ap _k =the AP of class , n= the number of classes

Dice Similarity Coefficient (DSC): The Dice Index displays a match between the expected and ground truth segmentation. This indicator considers both erroneous predictions and missing values in each class, therefore it not only counts the number of properly labeled pixels but also the segmentation borders' accuracy [21].

$$Dice = (2TP)/(2TP + FP + FN) \quad (3)$$

4.2. Benchmarking and Analysis

4.2.1. Detection result on Traffic sign

A comparison of the results on our dataset of the different versions of YOLOv5v6 is shown in table 3, the different models do not admit the same number of parameters ,it is clear presented that the one that admits more parameters is the best ,YOLOv5_l6 surpass YOLOv5_m6 by 2 % mAP. Compared to the passive learning strategy (see figure 5 (a, b, c, e) and figure 7 (a, b, c)), the model can correctly recognize any object class related to the traffic sign in both cases. On the other hand it failed to detect the textual data and this is normal because with an object detector remains limited in applying it to detect the text in scene. (see figure 5(d, f) and figure 7(d, e)).

4.2.2. Detection results on data of textual traffic

Testing our dataset on text detectors like CRAFT [22] and DB [23] shows a good performance (see figure 6) and both languages are detected perfectly.

4.2.3. Active learning vs. passive learning performance

Implementation details. With the totality of 3000 images collected with a definition of 1000*500 are divided between

Table 3: Comparison of the accuracy of various object detectors on NaSTSArLaT.

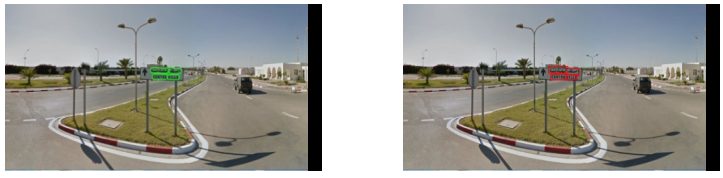
Models	Backbone	Size	Parameters	GFLOPs	mAP(%)
YOLOv5_s	CSPDarnket	640	7046599	15.9	44.1
YOLOv5_x	mobilenetv2	416	5958728	41.5	61.6
YOLOv5_m6	CSPDarknet	640	20907687	48.2	69.2
YOLOv5_l6	CSPDarknet	640	46186759	108.1	**69.4**

Table 4: Results on Arabic Text Detection Datasets.

Labelling	Training Data	Precision (%)	Recall (%)	mAP (%)	Nb of Annotation Data	
					Automatic	Manual
Session 1	181	34.8	54.9	45.6	31	169
Session 2	350	57.1	52.6	51.3	52	108
Session 3	458	59.1	52.8	53.2	45	55
Session 4	513	53.5	58	53	47	43
Session 5	556	73.2	50.3	56	131	69
Session 6	625	70.7	51	57.1	180	70
Session 7	695	70.9	53.4	61.4	219	81
Session 8	776	74.3	57.2	62.2	230	70
Session 9	846	83.5	55.9	**62.7**	232	68
Session 10	914	69.7	62.4	62.4	228	72
Session 11	986	72.9	58.6	62	226	74

Table 5: Active learning vs. passive learning performance

Annotation Method	Samples for Training	Time for Training(s)	mAP (%)	Processing time(s)	Dice index
PL_Annotation	13061	17725	69.2	36.4	0.40
AL_Annotation	3265	40268	62.7	40.6	0.67



(a) (b)

Fig. 6: A visual representation of the test results of different models on Data of textual traffic (a) DB, (b) CRAFT

training and validation data, included between 2700(90%),300 (10%) respectively.The object detection model is trained using two methods one with active learning and the other is arbitrary selected of data.In our experiment the version of the YOLOv5 model used is YOLOv5m6, the system is converged in the 9th iterations where the mean average remains stable to the previous one. The Google Colab platform with a 12GB NVIDIA Tesla K80 GPU is used for model training and assessment.

Deep Active Learning stratégie . table 4 compares the efficiency of the model for each iteration in terms of mAP until 9, its value is in progression, it becomes stable afterwards, so we don't need to continue, the mAP is around 62.7% and the quantity of training instances is 3165 which represents approximately 1/4 of the total of samples. In contrast, we acquire a value of 69.2% of mAP via random learning (see table 3).

Deep Active Learning results. According to the mAP of two types of training we made a comparison between the annotation approach using Active learning noted by (*_AL) and the other one known by random deep learning, see table 5.

Based on the mAP and Dice index metrics mentioned in table 5, we notice that with approximately 1/4 of the training samples we achieved the mAP of 62.7% as with the whole training instances.This good performance comes back to the fact that we found the right training data thanks to the successive iterations where each time the images having no object detected will be added with the previous amount of the base to form a new input of the new iteration . As it is shown in table 5 we can see that there is a big margin between the time provided by applying the active learning approach and the other one with passive learning, this is explained by the fact that the first one, the 3 CNN models are called at each iteration, on the other hand for the random approach the CNN model is

trained only once for the whole dataset. Based on these results, we can deduce that our technique avoids a laborious and time consuming task. (see figure 7)

Fig. 7: Visualization of the test results on traffic sign using deep active learning approach.

5. Conclusion

In this article we introduce a designated dataset for traffic sign and Arabic-Latin text using semi-Automatic process to facilitate the annotations of data, this dataset is available for the research community. We adopted the active learning approach after collecting the images using Google Street View. A Dataset for Arabic-Latin vocabulary road text is a challenging work because of the variety of types traffic sign and complexities background. A small amount of data is enough to determine a good performance in the detection of traffic sign and text in panel guide. On average, the quantity of training instances will be reduced to 1/4th of the original training size, according to the results. We have also demonstrated the effectiveness of NaST-SArLaT on evaluating it with the state of the art object/Text detection methods. The future work is to focus on the recognition of Arabic text existing in panel guide

References

[1] Peter A Hancock, Illah Nourbakhsh, and Jack Stewart. On the future of transportation in an era of automated and autonomous vehicles. *Proceedings of the National Academy of Sciences*, 116(16):7684–7691, 2019.

[2] Honggang Zhang, Kaili Zhao, Yi-Zhe Song, and Jun Guo. Text extraction from natural scene image: A survey. *Neurocomputing*, 122:310–323, 2013.

[3] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1454–1459. IEEE, 2017.

[4] Kai Wang and Serge Belongie. Word spotting in the wild. In *European conference on computer vision*, pages 591–604. Springer, 2010.

[5] Khalil Boukthir, Abdulrahman M Qahtani, Omar Almutiry, Habib Dhahri, and Adel M Alimi. Reduced annotation based on deep active learning for arabic text detection in natural scene images. *Pattern Recognition Letters*, 2022.

[6] Mohammed Akallouch, Kaoutar Sefrioui Boujemaa, Afaf Bouhoute, Khalid Fardousse, and Ismail Berrada. Asayar: A dataset for arabic-latin scene text localization in highway traffic panels. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[7] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011.

[8] Andreas Mogelmose, Mohan Manubhai Trivedi, and Thomas B Moeslund. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1484–1497, 2012.

[9] Alvaro Gonzalez, Luis M Bergasa, and J Javier Yebes. Text detection and recognition on traffic panels from street-level imagery using visual appearance. *IEEE Transactions on Intelligent Transportation Systems*, 15(1):228–238, 2013.

[10] Zhe Zhu, Dun Liang, Songhai Zhang, Xiaolei Huang, Baoli Li, and Shimin Hu. Traffic-sign detection and classification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2110–2118, 2016.

[11] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In *The 2013 international joint conference on neural networks (IJCNN)*, pages 1–8. Ieee, 2013.

[12] Xuejian Rong, Chucai Yi, and Yingli Tian. Recognizing text-based traffic guide panels with cascaded localization network. In *European Conference on Computer Vision*, pages 109–121. Springer, 2016.

[13] Maroua Tounsi, Ikram Moalla, and Adel M Alimi. Arasti: A database for arabic scene text recognition. In *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pages 140–144. IEEE, 2017.

[14] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1083–1090. IEEE, 2012.

[15] Saad Bin Ahmed, Saeeda Naz, Muhammad Imran Razzak, and Rubiyah Bte Yusof. A novel dataset for english-arabic scene text recognition (eastr)-42k and its evaluation using invariant feature extraction on detected extremal regions. *IEEE access*, 7:19801–19820, 2019.

[16] Rajae Moumen, Raddouane Chiheb, and Rdouan Faizi. Real-time arabic scene text detection using fully convolutional neural networks. *International Journal of Electrical and Computer Engineering*, 11(2):1634, 2021.

[17] Xiao Ke, Jiawei Zou, and Yuzhen Niu. End-to-end automatic image annotation based on deep cnn and multi-label data augmentation. *IEEE Transactions on Multimedia*, 21(8):2093–2106, 2019.

[18] Haiyu Song, Pengjie Wang, Jian Yun, Wei Li, Bo Xue, and Gang Wu. A weighted topic model learned from local semantic space for automatic image annotation. *IEEE Access*, 8:76411–76422, 2020.

[19] Oussama Zayene, Jean Hennebert, Sameh Masmoudi Touj, Rolf Ingold, and Najoua Essoukri Ben Amara. A dataset for arabic text detection, tracking and recognition in news videos-activ. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 996–1000. IEEE, 2015.

[20] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 399–407. Springer, 2017.

[21] Intisar Rizwan I Haque and Jeremiah Neubert. Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked*, 18:100297, 2020.

[22] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019.

[23] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11474–11481, 2020.