This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092821, IEEE Access

Digital Object Identifier

ATTICA: A dataset for Arabic Text-based Traffic Panels detection

Kaoutar SEFRIOUI BOUJEMAA[1], Mohammed AKALLOUCH[1], Ismail BERRADA[2], Khalid FARDOUSSE[1] and Afaf BOUHOUTE[1]
[1]Sidi Mohamed Ben Abdellah University (USMBA), Fez, Morocco.
[2]Mohammed VI Polytechnic University, School of Computer Science, Benguérir, Morocco.
Corresponding author: Kaoutar SEFRIOUI BOUJEMAA (kaoutar.sefriouiboujemaa@usmba.ac.ma).

ABSTRACT Detection and recognition of traffic panels and their textual information are important applications of advanced driving assistance systems (ADAS). They can significantly contribute in enhancing road safety by optimizing the driving experience. However, these tasks might face two major challenges. First, the lack of suitable and good quality datasets. Second, the in-feasibility of global standardization of traffic panels in terms of shape, color and language of the written text. Present research is intensively directed toward Latin text-based panels, while other scripts such as Arabic remain quiet undervalued. In this paper, we address this issue by introducing ATTICA[a], a new open-source multi-task dataset, consisting of two main sub-datasets: ATTICA_Sign for traffic signs/panels detection and ATTICA_Text for Arabic text extraction/recognition. Our dataset gathers 1215 images with 3173 traffic panel boxes, 870 traffic sign boxes and 7293 Arabic text boxes. In order to examine the utility and advantages of our dataset, we adopt state-of-the-art deep learning based approaches. The conducted experiments show promising results confirming the valuable addition of our dataset in this field of research.

[a]https://github.com/kkawtar/ATTICA

INDEX TERMS Traffic panels, sign detection, sign recognition, scene Arabic text detection, traffic textual information retrieval, traffic panels dataset.

I. INTRODUCTION

OVER the past decade, AI (Artificial Intelligence) has remarkably contributed to the development of innovative technologies in numerous sectors (e.g. health, transportation, education, and robotics) [1]–[3]. In the transportation sector, AI is widely used for developing driving assistance and automated driving solutions. Advanced Driving Assistance Systems (ADAS) [4] are life-saving technologies [5], designed to offer different driving assistance features using all sources of traffic data (e.g. automatic emergency braking, driver's distraction warning, speed adaptation and traffic signs detection [6]). In particular, computer vision is one of the key technologies that have helped accelerating the evolution of ADAS and their real-time performance.

From an AI perspective [7], the key limiting factor to overpass AI challenges is the availability of high quality data. In fact, all AI accomplishments that are near human level performance have only been possible with the release of appropriately designed data. The 2014 Google's breakthrough in the field of image classification is due to a new image object classifier called GoogleNet trained on the ImageNet corpus [8]. Since then, new convolutional neural network (CNN) architectures have been developed and researchers/industries have started to realise the importance of collecting and annotating quality data [9].

Detection of traffic signs and their textual information are important tasks for ADAS [10]. Their main value is to ensure the drivers' safety by preventing them from ignoring mandatory traffic information (e.g. speed limit) or from being distracted when reading signs while driving. Having this kind of assistance can definitely contribute in reducing the frequency of severe traffic accidents. Traffic signs detection can additionally be employed for building automatic visual inspection systems of signs and panels, for inventory and maintenance purposes [11]. Unfortunately, computer vision researchers still face major data challenges when addressing these tasks. First, most of the available datasets include only three types of traffic signs: regulatory, mandatory and warning, whereas traffic panels (a.k.a guide panels) are rarely available. Yet, route information (Directions, places

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092821, IEEE Access

names, mileage information, main road name, etc) is only displayed on guide panels. Second, datasets for scene text localization and recognition are often limited to Latin script, which is considerably easier to process compared to cursive scripts (e.g. Japanese, Chinese, Persian and Arabic) [12], [13]. Moreover, these datasets are occasionally collected in a transportation context, which is again discouraging for road safety researchers.

In this paper, we aim to address the lack of traffic panel datasets with Arabic scripts, by introducing ATTICA, a new multi-task dataset of traffic signs/panels. ATTICA is a challenging dataset since it gathers real-world visual complex scenes that are captured in challenging traffic environments. More precisely, ATTICA includes samples with low resolution, complex background, noise, different text alignments and variations in terms of size, color and style. The collected dataset consists of two major sub-datasets:

1) ATTICA_Sign: contains annotations of different types of traffic signs/panels objects.
2) ATTICA_Text: contains text objects with line and word level annotations.

ATTICA contains 1215 images with 3173 traffic panel boxes, 870 traffic sign boxes and 7293 Arabic text boxes, collected from open-source images on the internet. We have adopted a careful manual annotation for the different boxes using 7 object categories, namely, traffic panel, traffic sign, other sign, km-point, add-panel, text line-level (including 2 sub-categories), text word-level (including 5 sub-categories).

In order to illustrate the usability of the introduced dataset, we conduct two major experiments. The first experiment is related to the traffic panels/signs detection task, in which we evaluate and compare the performance of four well-known CNNs-based architectures: Single Shot multibox Detector (SSD) [14], Region-based Fully Convolution Network (R-FCN) [15], Faster-RCNN [16] and RetinaNet [17]. In the second experiment, we tackle the task of Arabic text line detection in traffic panels/signs by adopting the famous CTPN (Connectionist Text Proposal Network) and EAST (Efficient and Accurate Scene Text) models [18], [19]. Our findings show promising results which definitely validate the quality of our dataset.

To the best of our knowledge, this is the first traffic text-based panels dataset that is collected from the internet and which contains data from multiple Arab countries (e.g. Morocco, Algeria, Egypt, Saudi Arabia and many others). This feature allows new investigations for building standardized traffic panel detectors to be applied in Arabic regions.

The rest of this paper is organized as follows. Section II reviews existing datasets for traffic sign and text scene detection. A detailed description of the introduced dataset is highlighted in section III. The adopted approaches and evaluation metrics are presented in section IV. Experiments and findings are discussed in section V. Finally, we conclude by summarizing our contributions and future directions in section VI.

II. RELATED WORKS

In this section, we discuss some of the well-known open-source datasets, used for evaluating Traffic Sign (TS) and text-based Traffic Panel (TP) detection.

A. TRAFFIC SIGNS DATASETS

Tsinghua-Tencent 100k is one of the recently published TS datasets [20]. It contains over 100k images, collected from the Tencent Street Views of 300 Chinese cities. The dataset provides 30k TS boxes annotated using pixel masks and bounding boxes (bboxes). The Russian TS Dataset (RTSD) is another considerably large set, having 179138 labelled frames with 156 sign categories [21]. RTSD provides 104358 TS annotations, surpassing the previously mentioned dataset. LISA is a TS recognition dataset that includes different videos recorded in the United States [22]. It has 6610 frames with 7855 annotations divided into 47 sign categories. The Swedish TS detection and recognition database is a collection of 20k frames recorded over 350km of the Swedish highways and city roads [23]. It provides 3488 TS boxes that are annotated using box coordinates, sign type, visibility status and road status. Unfortunately, these annotations only represent 20% of the data. The German TS Recognition Benchmark (GTSRB) is one of the highly known and challenging TS datasets [24]. It is also considered the first to help evaluate significantly the problem of TS classification. GTSRB contains 50k images annotated using 43 sign categories. Following that, the German TS Detection Benchmark (GTSDB) was introduced [25]. It is a collection of 900 images captured on the German roads. Signs are labeled using only 3 categories (mandatory, danger and prohibitory). The Belgium TS database is quite similar to the German sets [26]. It includes two major TS datasets for detection and recognition, with over 13k annotations of 145k images taken on Belgian roads. Signs are annotated using 63 categories for recognition and 3 categories for detection as in GTSDB.

B. TEXT-BASED TRAFFIC PANELS DATASETS

In the literature, the commonly suggested approach for TP (Traffic Panel) detection is based on color and shape segmentation techniques [27], [28]. Unfortunately, these latter do not meet the efficiency requirements of real-time applications.

Accordingly, the hypothesised methodology for TP's text extraction consists of training a text detector/recognizer on an outdoor scene text dataset. These sets are not necessarily collected in a road context and focus heavily on Latin scripts [29], [30]. COCO-text [31], ICDAR [32]–[34] and SVT [35] are some of the widely used Latin datasets. They contain mainly English text collected from various sources such as announcements, books and posters. As for Arabic scripts, there exist only few benchmarking datasets such as ALIF [36] and AcTiv [37] that are collected from Arabic news channels. The mentioned sets can be quite inconvenient for TP text applications since they are annotated on line and word levels which do not correspond with the route text vocabulary (a.k.a, words bag). ASAYAR dataset has been recently pub-

TABLE 1. Open source Datasets compared to our dataset.

Detection Dataset		COCO-text	ICDAR	SVT	ALIF	AcTiV	GTSRB	LISA	Tsinghua-Tencent 100K	RTSD	ASAYAR	ATTICA
Traffic Sign							✓	✓	✓	✓	✓	✓
Traffic Panel											✓	✓
Other-Sign (see Fig. 4(b))												✓
Km-Point (see Fig. 4(c))												✓
Add-Panel (see Fig. 5)												✓
Arabic Text	**Word-level**				✓						✓	✓
	Line-level				✓	✓					✓	✓
Data collection from multiple countries												✓

lished to especially address this issue [38]. ASAYAR includes 1763 images collected on different Moroccan highways and annotated using 16 object categories. This set provides over 19k annotated boxes in which TP and Arabic/French text related boxes are prioritised. Another similar dataset was introduced in [39], focusing on TP in urban areas. The set contains 26988 frames captured on Iranian roads, in which 5040 are annotated as TP and Persian/English text bboxes.

To summarize, a detailed comparison of the above mentioned datasets is listed in Table 1. Compared to these sets, as shown in Section III, ATTICA is a challenging and good quality data source for text-based and symbolic TS/TP research. In addition, ATTICA introduces new TS categories not investigated in previous studies.

FIGURE 1. Samples of images taken under various periods of the day and weather conditions. (a) Sunny weather. (b) Evening. (c) Night. (d) Normal weather. (e) Cloudy and rainy. (f) Mirage vision caused by hot weather.

III. DATASET CONSTRUCTION

In this section, we review important details of the ATTICA dataset, including data collection, data annotation and a description of ATTICA_Sign and ATTICA_Text sub-datasets.

A. DATA COLLECTION

ATTICA is a collection of 1215 images representing captured roadway scenes with various types of traffic signs (TS) and traffic panels (TP), collected from open-source images on the internet. This diversity is carefully achieved by considering three main conditions in the collection process:

1) Selecting scenes that are captured in day and night times, as well as in various weather conditions, in order to cover most of the visual challenges that can directly impact the detection task (see Fig. 1).
2) Selecting scene texts from two Arabic regions namely, North Africa (Algeria, Egypt, Morocco, and Tunisia) and Gulf (Bahrain, Kuwait, Qatar, Saudi Arabia, and the United Arab Emirates) regions (see Fig. 2)
3) Selecting scenes of different roadway types, including city roads, national roads and highways, with the objective of covering most of the related TS/TP variations such as shape, position, background complexity, color and textual content (see Fig. 3 and Fig. 4).

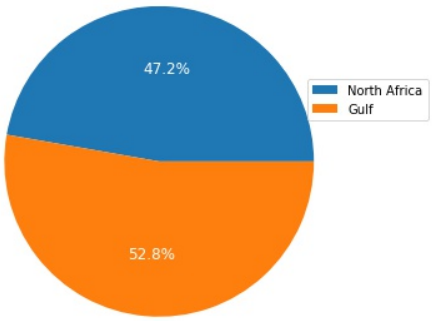

FIGURE 2. Distribution of the ATTICA dataset samples according to North Africa and Gulf regions.

B. ANNOTATION

The annotation process of the ATTICA dataset was carried out during a period of 10 months, by a group of four researchers using Labelme tool. Labelme automatically generates XML metadata files according to the Pascal VOC format (the image name, size, object bboxes coordinates and corresponding class names). In addition, a CSV file is provided to indicate the downloadable source links of all images, along with other information describing the contained traffic panels (color, shape, location, type and noise presence). We divide our dataset into two main sub-datasets:

- ATTICA_Sign: contains annotations of different types of traffic signs/panels objects.
- ATTICA_Text: contains text objects with line and word level annotations.

As reported in Table 2, ATTICA contains a total of 1215 images, 1180 of them include text-based signs/panels. In

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092821, IEEE Access

TABLE 2. Composition of the ATTICA dataset.

Sub-dataset	Nb of images	Nb of classes
TS/TP	1215	5
Text	1180	4

what follows, the structure of the two sub-datasets is detailed.

1) ATTICA_Sign

We consider five categories for annotating traffic signs objects in our dataset:

- **Traffic Panels (TPs):** big-size guiding signs used to provide directional and mileage information. TPs can have various designs depending on their use cases. For example, on highways, TPs usually take rectangular or arrow shapes and vary between three colors: blue, white and green (see Fig. 3(a) and Fig. 3(b)). Another type of TPs which is neglected in the state-of-the-art, is dynamic panels (see Fig. 3(c)). They are used for displaying information about varying traffic conditions (e.g. traffic delays, warning messages and others).

FIGURE 3. Samples of different types of Traffic Panels. (a) Rectangular shaped traffic panel. (b) Arrow shaped traffic panels. (c) Dynamic traffic panel for varying traffic conditions display.

- **Traffic Signs (TSs):** small signs that can be found at the side of roads. TSs are used for indicating regulatory, mandatory and warning instructions to road users (e.g. speed limit, stop, pedestrian crossing, warning of road works, etc). Unlike TPs, TSs include limited text content that is usually based on symbols and numbers. In addition, TSs are generally designed using three forms: circle, triangle and octagon, and three colors: blue, red and yellow (see Fig. 4(a)).
- **Other-Sign (OS):** warning signs used to indicate route sharp left/right deviations, roundabouts deviation, end of a bridge parapet, tunnel entrance, traffic cones and roads barriers. Usually, these signs do not include any text and their design is mostly restricted to thin rectangular shapes with sloping bars or chevrons (see Fig. 4(b)).
- **Km-Point (KP):** road signs made of rocks, which are generally placed at the side of national roads. This type of road signs is mainly used to include place names and mileage information (see Fig. 4(c)).

FIGURE 4. Samples of (a) Traffic Signs, (b) Other-Signs and (c) Km-Points.

- **Add-Panel (AP):** special plates added to TPs/TSs or placed at the side of highways for providing important supplementary information. In case of TPs (see Fig. 5(a)), APs generally include the main route ID, whereas for TSs (see Fig. 5(b)), they include information indicating when the sign's instructions will be valid or which road users are affected by the sign, etc. However, an AP placed at the side of a highway is only used to include mileage information (see Fig. 5(c)).

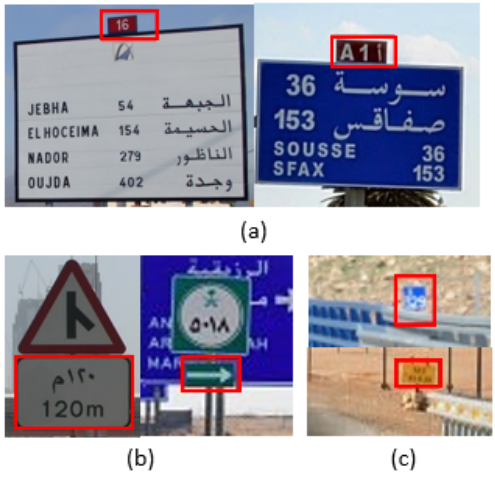

FIGURE 5. Samples of different types of Add-Panels. (a) Add-panels joined to Traffic Panels. (b) Add-Panels joined to Traffic Signs. (c) Add-Panels on sides of highways.

ATTICA_Sign is labeled according to the aforementioned classification. As stated in Table 3, it contains 4607 bboxes objects, where ∼69% belongs to the "TP" category and ∼18% to the "TS" category (see Fig. 6). This distribution

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092821, IEEE Access

TABLE 3. ATTICA_Sign class distribution.

Classes	Nb of boxes	Average by image
Traffic Panel	3168	2.61
Traffic Sign	822	0.68
Other-Sign	406	0.33
Add-Panel	196	0.16
Km-Point	15	0.01

reflects an important class imbalance that needs to be considered in the training process.

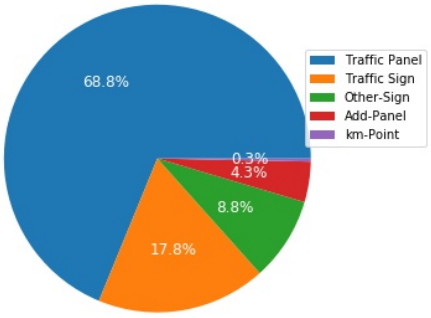

FIGURE 6. Distribution of ATTICA_Sign categories.

2) ATTICA_Text

Almost all ATTICA_Sign categories are composed of Arabic text. To generate quality line and word level text detection and recognition data, the following categories are considered:

- **Line-level categories:**
  - -- Readable line: Arabic text lines that can be clearly viewed and read (see Fig. 7(a)).
  - -- Unreadable line: Arabic text lines that are difficult to read. To our best of knowledge, such annotation has never been proposed before for outdoor scene detection. However, this category can have multiple interesting applications such as detecting damaged TPs that need maintenance, or re-positioning TPs to enhance their visibility for drivers, etc. In addition, this category can be useful for accelerating and enhancing training of a line-level text detector (see Fig. 7(b) and Fig. 7(c)).

FIGURE 7. Samples of Line-level categories. (a) Readable line. (b) and (c) Unreadable lines.

- **Word-level categories:**
  - -- Arabic word (see Fig. 8(a)).
  - -- Arabic digit (see Fig. 8(b)).
  - -- Special character (see Fig. 8(c)).
  - -- Latin digit (see Fig. 8(d)).
  - -- Latin unit: Latin mileage unit text. In some cases, it is only written in Latin even if all panel information is in Arabic. Therefore, our decision for considering this category is mainly to support semantic analysis studies (see Fig. 8(e)).

The distribution of the above mentioned categories is listed in Tables 4 and 5, and visualized in accordance with ATTICA_Sign in Fig. 9. These statistics reveal an overall interesting number of annotated text bboxes, where 14570 bboxes are for word-level detection and recognition. This proves the significant utility of our dataset for building robust outdoor scene text detectors, especially in traffic environments.

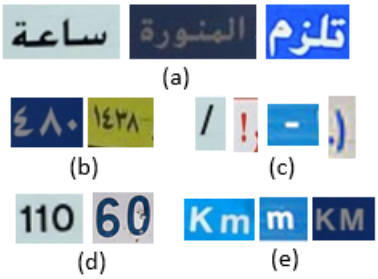

FIGURE 8. Samples of Word-level categories. (a) Arabic word. (b) Arabic digit. (c) Special character. (d) Latin digit. (e) Latin unit.

TABLE 4. ATTICA_Text Line and Word levels distribution.

Class	Nb of boxes	Average by image
Line-level	7293	6.21
Word-level	14570	12.34

TABLE 5. ATTICA_Text categories distribution.

LINE-LEVEL		
Class	Nb of boxes	Average by image
Readable line	6457	5.5
Unreadable line	836	0.71
WORD-LEVEL		
Arabic word	11494	9.74
Special character	473	0.4
Arabic digit	628	0.53
Latin digit	1503	1.27
Latin unit	472	0.41

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092821, IEEE Access

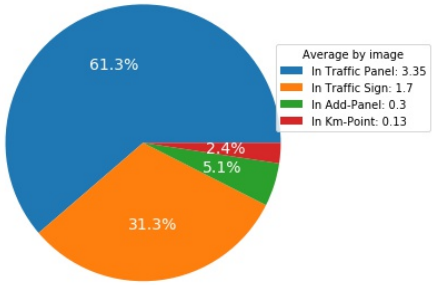

FIGURE 9. Distribution of the "Readable-line" category according to ATTICA_Sign.

C. CHALLENGES

The ATTICA dataset has various substantial challenges in terms of scene object detection. As we previously mentioned, our data come from different Arabic countries and are captured under different environmental conditions. Thus, there is a wide diversity in the annotated objects for both ATTICA_Sign and ATTICA_Text sub-datasets. In particular, road signs can be challenging to process mainly for their position/condition, shape, color, content density, camera angle and lightning. However, for text, there are some other exceptions such as text alignment, noisy backgrounds, font style, size and color. Note that, having the cursive nature of the Arabic script is enough competitive for processing. Class imbalance is also considered as a challenging aspect for ATTICA_Sign, since $\sim$69% of the data belongs to the TP category and only 7.5% to the AP and KP categories.

Nevertheless, we consider these challenges as quality data features since they can contribute in the enhancement of building real-world robust object detectors.

IV. BASELINES

In order to evaluate the utility of the ATTICA dataset, we adopt a baseline of state-of-the-art Neural Networks "NN" architectures for object detection and recognition, as well as text detection.

A. OBJECT DETECTION AND RECOGNITION

1) Faster R-CNN [16]

was first introduced in 2015, as an improved version of Fast R-CNN [40]. It is a two-stage Convolutional Neural Network (CNN), used for scene object detection and classification. The first stage is achieved through a CNN base model and a Region Proposal Network (RPN). The CNN base model is mainly composed of convolution, activation and pooling layers. They are responsible for generating a features map F for an input image I. Generally, the architecture of the base model is inspired by recognized CNN image classifier, such as VGG, AlexNet and ResNet [41]–[43]. The RPN, on the other hand, takes the feature map F as input and generates object proposals, called "Regions Of Interest" ROIs. RPN is simply composed of 3 convolution layers. The first layer performs two essential tasks: (1) generating a fixed number of ROIs having different shapes, by sliding over each location in F and (2) assigning positive and negative labels to the proposed ROIs, based on their Intersection over Union (IOU) value with the ground truth boxes. The second and third layers are Fully Connected (FC). They are used for parallel box object classification and bbox regression, for a set of K ROIs. The classification layer outputs a binary 2K-d vector, indicating 0 for no-object and 1 for object. As for the regression layer, it outputs a 4K–d vector, indicating the new adjustment of the ROIs bboxes coordinates $\{x_{min}, y_{min}, x_{max}, y_{max}\}$. Note that, both layers use ground-truth boxes categories and bounding coordinates for their computations. Finally, the second stage of the Faster R-CNN, is composed of (1) pooling layers for reshaping the generated ROIs and (2) FC layers to perform per-category object classification and final regression to adjust the ROIs bboxes coordinates.

2) Single Shot multibox Detector (SSD) [14]

is a one–stage CNN detector, first introduced in 2016. It is inspired by the Faster R-CNN architecture. SSD is composed of (1) a CNN base model for deep features extraction, (2) auxiliary $n \times n$ convolution layers (with activation layers) that progressively shrink in size, to generate multiple feature maps $\{F\}$ of variable sizes, (3) convolution layers to generate, for each $f_i \in \{F_i\}$, ROIs proposals matching the ground-truth boxes categories and (4) FC layers, to compute final object category classification and bbox regression. Finally, a Non-Maximum Suppression (NMS) layer is employed to select the top N likely predictions, using confidence and IOU thresholds. Note that, the class imbalance issue is interestingly addressed in SSD, via the "hard negative sampling" technique. This latter consists of maintaining a predefined ratio of positive and negative samples (e.g. 1:3) during the loss function computation.

3) Region-based Fully Convolution Network (R-FCN) [15]

is a two-stage CNN detector, introduced in 2016. Its architecture is inspired by Faster R-CNN and FCN [44]. R-FCN introduced new concepts called “position sensitive score maps” and “voting maps” [45]. These two concepts mainly contribute in speeding the process of objects detection and classification. They additionally help increasing the model’s accuracy. The first stage of R-FCN consists of a CNN base model for generating a feature map F or different feature maps $\{F_i\}$ in case of multi-scale training (cf. SSD). At the second stage, F is fed to two parallel networks: (1) an RPN for generating ROIs proposals (without FC layers) and (2) two $n \times n$ convolution layers for computing $k^2(c+1)$ score maps S for object category classification and $4k^2 - d$ vectors D for bboxes coordinates regression. Moreover, R-FCN applies for each ROI, a set of $n \times n$ position sensitive pooling layers to create 1) a classification voting map V_1 that indicates the similarity likelihood of the ROI and its corresponding object category in S and 2) a bbox regression voting map V_2 following the same concept with D. Finally,

these voting maps are passed to FC layers for final object category classification and bbox regression. R-FCN adopts the same total loss function of Faster R-CNN.

4) RetinaNet [17]

is a one-stage detector, introduced in 2018. Its architecture is composed of the Feature Pyramid Network (FPN) [46] as a base model (built on top of ResNet), for generating multi-stage feature maps $\{F_i\}$. Each map is connected to a FCN, composed of two parallel sub-networks for final object category classification and bbox regression. Note that, FPN allows creating multi-scale feature maps (of high and low resolutions) with strong semantic levels, which leads to finite detection results. RetinaNet is capable of achieving the speed of one-stage detectors and over-passing the accuracy of two-stage detectors. This is supported by the use of the so called "Focal loss" function which addresses the issue of positive/negative class imbalance by highly penalizing hard negative samples.

B. TEXT DETECTION

1) The Connectionist Text Proposal Network (CTPN)

is an end-to-end trainable NN, introduced by Z.Tian et al in 2016 [47]. It is designed to detect text lines in natural scene images by adopting the characteristics of CNNs. Notably, CTPN can handle multi-scale and multi-lingual text without the need of any post-processing. There are three main parts composing the CTPN architecture. First, a VGG16 base model for deep feature map extraction F. Second, a recurrent in-network for $n \times n$ window sliding over F. The generated text proposals are then passed to a Bi-directional LSTM network [48] for sequence connecting. Finally, a 512D FC layer is adopted for final bboxes classification (text/non text) and regression. Noting that, outputs of CTPN are sequential fixed-width and fine-scale text proposals.

2) Efficient and Accurate Scene Text Detector (EAST)

is a fully CNN, first introduced by Zhou et al in 2017 [19]. It is designed for word and line level scene text detection. The architecture of EAST is composed of three branches combining a single NN. First, a "Feature extractor stem" is placed for multi-scale deep feature maps $\{F_i\}$ extraction. For this step, a combination of VGG16 and PVANET [49] as a base model is adopted. Second, a "Feature merging branch" is included as a U-shape network composed of convolution and unpooling layers for gradual feature maps ($\{F_i\}$) merging into a single map F. The final branch, called "Output layer" is a set of FC layers for text/non-text classification and geometry (bbox axis alignment and rotation angle) regression.

V. EXPERIMENTS AND RESULTS

In this section, we demonstrate the usability of the ATTICA dataset for sign and text line level detection. For this end, we evaluate the performance of a range of state-of-the-art NNs architectures. These last are presented in Section IV.

A. EXPERIMENTAL SETUP

1) Train and Test sets

we adopt a stratified sampling technique to split data into train (80%) and test (20%) sets. The number of images in the (train, test) sets for ATTICA_Sign and ATTICA_Text are (960, 240) and (944, 236), respectively. Note that the "Km-Point" category is excluded from the ATTICA_Sign experiments, and we only consider samples of readable lines.

2) Implementation framework

The baseline models are implemented using TensorFlow and pre-trained object classifiers as backbones (Table 6). Given the wide range of object scales in our dataset, various default boxes scales (see Table 7) and aspect ratios ({2:1, 1:1, 1:2}) are considered in training.

TABLE 6. Backbones of the experimented models.

Model	Backbone	Backend	Data
Faster R-CNN	ResNet101	Tensorflow	ATTICA_Sign
SSD	MobileNet		
R-FCN	ResNet101		
RetinaNet	ResNet50		
CTPN	VGG16	Tensorflow	ATTICA_Text (Line level)
EAST			

TABLE 7. Default boxes (a.k.a anchors) scales.

Model	Min	Max
Faster RCNN	128^2	512^2
SDD	0.2×300^2	0.9×300^2
R-FCN	300^2	1024^2
RetinaNet	32^2	512^2

3) Execution environment

Model training and evaluation are conducted on the Google Colab plateform with a 12GB NVIDIA Tesla K80 GPU.

B. EVALUATION METRICS

For both Sign and Text detection, the following state-of-the-art metrics are adopted.

1) Intersection Over Union (IoU)

measures the goodness of fit of a predicted bbox B_p with respect to a ground truth bbox B_t. IOU computes the overlapping area ratio of both bboxes as follows:

$$IoU = \frac{area(B_p \cap B_t)}{area(B_p \cup B_t)} \quad (1)$$

Given a threshold $\alpha \in [0, 1]$ (e.g, $\alpha = 0.7$), B_p is classified as a True Positive (TP) only if $IoU \geq \alpha$, otherwise it is a False Positive (FP). False Negatives (FN) are accumulated if the model has no output predictions for an image.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092821, IEEE Access

2) Average Precision (AP) & mean AP (mAP)

AP is a popular metric for measuring the accuracy of a detection model using Precision (p) and Recall (r) (see Eqs (2) and (3)). AP is a weighted sum of p values at each threshold, where the weight is the increase in r.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$AP = \sum_{i=0}^{N-1} \underset{r':r' \geq r_{(i+1)}}{max} p(r')(r_{(i+1)} - r_i) \tag{4}$$

mAP is the average of AP values for all objects in a detection model. mAP is used to encapsulate the accuracy of the overall predictions (bbox and category classification) in one value.

C. BENCHMARKING AND ANALYSIS

1) ATTICA_Sign results

The benchmarking results of state-of-the-art models are summarized in Table 8, and example of images showing the detection results are displayed in Fig. 10. In general, we notice a remarkable out-performance of R-FCN, achieving very interesting AP scores $\in [0.87, 0.99]$. The SSD model, ranked as second, showing quality results with similar performance for the Traffic Panel class. These significant achievements of both R-FCN and SSD models can be due to their characteristic of multi-scale feature maps generation. On the other hand, results of the Faster R-CNN and RetinaNet are surprisingly in line. In spite of their good AP scores (0.68 and 0.76, respectively) for the dominant class (Traffic Panel), they show poor performance in detecting other classes (AP scores $\in [0.23, 0.31]$). In addition to the severe class imbalance, these poor results can be explained by the high amount of very small object boxes in the Traffic Sign, Other-Sign and Add-Panels categories.

To further investigate the performance of these two models (Faster R-CNN and RetinaNet), we conduct a second experiment where we only focus on dominant classes (Traffic Panel and Traffic Sign categories). The two models are trained and evaluated on the ATTICA_Sign, where only object boxes of scales $\geq 16^2$ and $\geq 32^2$ are selected. The obtained results are summarized in Table 9, demonstrating, as previously discussed, better AP scores especially for the Traffic Sign class (AP score $\in [0.58, 0.95]$).

2) ATTICA_Text results

The results for line level detection using CTPN and EAST models are highlighted in table 10. We adopt Precision and Recall as metrics for performance evaluation. Analyzing the obtained results, we notice a similar behaviour for both models; Recall scores are > 0.66 and Precision scores are < 0.5. These scores can be explained by some characteristics and challenges presented in the ATTICA dataset. First, there is a considerable heterogeneity of the text boxes, namely background color, text font style, color, position and noise in form of text, making the detection more challenging. Second, the "unreadable line" category data may add some confusion for the trained models, when some of its boxes are detected as text while they were excluded from both train and test data. These boxes are then classified as false positives which contributes in reducing Precision and respectively increasing Recall.

Data augmentation is one of the well known techniques for

TABLE 8. ATTICA_Sign results.

Model	Class	AP	mAP@0.5
Faster R-CNN	Traffic Panel	0.68	0.36
	Traffic Sign	0.25	
	Other-Sign	0.23	
	Add-Panel	0.27	
SSD	Traffic Panel	0.91	0.59
	Traffic Sign	0.40	
	Other-Sign	0.48	
	Add-Panel	0.56	
R-FCN	Traffic Panel	0.99	0.95
	Traffic Sign	0.94	
	Other-Sign	0.98	
	Add-Panel	0.87	
RetinaNet	Traffic Panel	0.76	0.40
	Traffic Sign	0.28	
	Other-Sign	0.24	
	Add-Panel	0.31	

TABLE 9. ATTICA_Sign results using Faster R-CNN and RetinaNet, by considering data samples with boxes of sizes $\geq 16^2$ and $\geq 32^2$.

		Boxes size			
		$\geq 16^2$		$\geq 32^2$	
Model	Class	AP	mAP	AP	mAP
Faster R-CNN	Traffic Panel	0.80	0.69	0.84	0.74
	Traffic Sign	0.58		0.63	
RetinaNet	Traffic Panel	0.79	0.83	0.81	0.88
	Traffic Sign	0.86		0.95	

TABLE 10. ATTICA_Text (Line level) results using CTPN and EAST models. For data, only readable text line levels are considered.

Model	Used data	Precision	Recall
CTPN	Line level	0.41	0.69
EAST	Line level	0.48	0.66

improving the overall model performance. To evaluate this, we conduct a second experiment, where the ASAYAR_TXT dataset is added [38]. The text line boxes in this dataset are noticeably homogeneous when compared to ours. ASAYAR_TXT includes about 1375 images, having 2165 Arabic text line box. Thus, the updated sizes of the (train, test) sets are (2044, 511). Results are exhibited in Table 11. In comparison with the previous experiment, we notice a remarkable enhancement in term of Precision scores ($>= 0.67$) for both CTPN and EAST. The quality of the ASAYAR_TXT in terms of noise and text boxes characteristics have definitely contributed in ameliorating the models performance. Detection results are displayed in Fig. 11.

TABLE 11. CTPN and EAST results for Text line level detection using our Text-line sub-dataset and the ASAYAR_TXT dataset.

Models	Used data	Precision	Recall
CTPN	Line level	0.67	0.85
EAST	+ASAYAR_TXT (Line level)	0.71	0.89

VI. CONCLUSION

In this paper, we have introduced ATTICA, a new open-source dataset for Arabic text-based traffic signs/panels detection. The dataset is publicly available and accessible for the research community. ATTICA is highly diverse and challenging, since it contains roadway scenes from various Arab countries. New traffic signs and text objects annotations are introduced in the ATTICA dataset, to allow further investigations. The methodology adopted for collecting and annotating our dataset, was carefully presented. In addition, the conducted experiments using state-of-the-art models for sign and text detection, demonstrate the quality of ATTICA. Finally, the performed experiments provided comprehensive results, indicating the possibility of real-time applications.

REFERENCES

[1] González García, C., Núñez Valdéz, E. R., García Díaz, V., Pelayo García-Bustelo, B. C., & Cueva Lovelle, J. M. (2019). A review of artificial intelligence in the Internet of Things. International Journal of Interactive Multimedia and Artificial Intelligence.

[2] Aghion, P., Jones, B. F., & Jones, C. I. (2017). Artificial intelligence and economic growth (No. w23928). National Bureau of Economic Research.

[3] Nayak, A., & Dutta, K. (2017, June). Impacts of machine learning and artificial intelligence on mankind. In 2017 International Conference on Intelligent Computing and Control (I2C2) (pp. 1-3). IEEE.

[4] Jiménez, F., Naranjo, J. E., Anaya, J. J., García, F., Ponz, A., & Armingol, J. M. (2016). Advanced driver assistance system for road environments to improve safety and efficiency. Transportation research procedia, 14, 2245-2254.

[5] Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. A. (2019). Applications of artificial intelligence in transport: An overview. Sustainability, 11(1), 189.

[6] Chan, C. Y. (2017). Advancements, prospects, and impacts of automated driving systems. International journal of transportation science and technology, 6(3), 208-216.

[7] Alexender Wissner-Gross. Datasets Over Algorithms. 2016 : What do you consider the most interesting recent [scientific] news? What makes it important?. https://www.edge.org/response-detail/26587.

[8] SZEGEDY, Christian, LIU, Wei, JIA, Yangqing, et al. Going deeper with convolutions. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 1-9.

[9] Zhang, Q., Zhang, M., Chen, T., Sun, Z., Ma, Y., & Yu, B. (2019). Recent advances in convolutional neural network acceleration. Neurocomputing, 323, 37-51.

[10] Khan, S., Thainimit, S., Kumazawa, I., & Marukatat, S. (2017, May). Text detection and recognition on traffic panel in roadside imagery. In 2017 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES) (pp. 1-6). IEEE.

[11] González, Á., Bergasa, L. M., Gavilán, M., Sotelo, M. A., Herranz, F., & Fernández, C. (2009, October). Automatic information extraction of traffic panels based on computer vision. In 2009 12th International IEEE Conference on Intelligent Transportation Systems (pp. 1-6). IEEE.

[12] CHANDIO, Asghar Ali, ASIKUZZAMAN, Md, PICKERING, Mark, et al. Cursive-text: A comprehensive dataset for end-to-end Urdu text recognition in natural scene images. Data in Brief, 2020, vol. 31, p. 105749.

[13] BHUNIA, Ankan Kumar, KONWER, Aishik, BHUNIA, Ayan Kumar, et al. Script identification in natural scene image and video frames using an attention based Convolutional-LSTM network. Pattern Recognition, 2019, vol. 85, p. 172-184.

[14] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[15] Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. In Advances in neural information processing systems (pp. 379-387).

[16] REN, Shaoqing, HE, Kaiming, GIRSHICK, Ross, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 2016, vol. 39, no 6, p. 1137-1149.

[17] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[18] Tian, Z., Huang, W., He, T., He, P., & Qiao, Y. (2016, October). Detecting text in natural image with connectionist text proposal network. In European conference on computer vision (pp. 56-72). Springer, Cham.

[19] ZHOU, Xinyu, YAO, Cong, WEN, He, et al. East: an efficient and accurate scene text detector. In : Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017. p. 5551-5560.

[20] Z.Zhu, D.Liang, S.Zhang, X.Huang. "Traffic-sign detection and classification in the wild". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016b) 2110–2118.

[21] Chigorin, A., & Konushin, A. (2013). A system for large-scale automatic traffic sign recognition and mapping. CMRT13–City Models, Roads and Traffic, 2013, 13-17.

[22] Andreas Møgelmose, Mohan M. 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, 2012.

[23] Larsson, F., Felsberg, M., & Forssen, P. E. (2011). Correlating Fourier descriptors of local patches for road sign recognition. IET Computer Vision, 5(4), 244-254.

[24] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. "The German Traffic Sign Recognition Benchmark: A multi-class classification competition". In Proc. IJCNN, pages 1453–1460, July 2011.

[25] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German traffic sign detection benchmark," in IJCNN. Ieee, aug 2013, pp. 1–8. 1, 2, 3.

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

[27] GONZALEZ, Alvaro, BERGASA, Luis M., et YEBES, J. Javier. Text detection and recognition on traffic panels from street-level imagery using visual appearance. IEEE Transactions on Intelligent Transportation Systems, 2013, vol. 15, no 1, p. 228-238.

[28] Guo, J., You, R., Feng, C., & Huang, L. (2018, August). Detection of street-level traffic panels based on cascaded color segmentation. In 2018 13th International Conference on Computer Science & Education (ICCSE) (pp. 1-6). IEEE.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3092821, IEEE Access

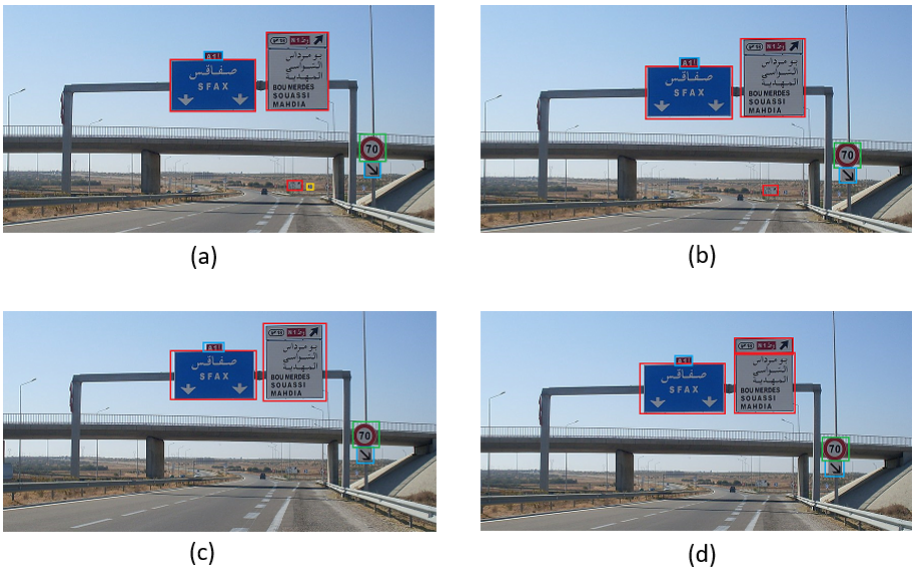

FIGURE 10. Visualization of the test results of different models (see Table 8) on ATTICA_Sign: (a) R-FCN, (b) SSD, (c) RetinaNet, (d) Faster R-CNN. Red: Traffic Panels, blue: Add-Panel, green: Traffic-Sign and yellow: Other-Sign

FIGURE 11. Visualization of the test results of different models (see Table 11) on ATTICA_Text (Line level): (a) EAST, (b) CTPN.

[29] Gonzalez, A., Bergasa, L. M., & Yebes, J. J. (2013). Text detection and recognition on traffic panels from street-level imagery using visual appearance. IEEE Transactions on Intelligent Transportation Systems, 15(1), 228-238.

[30] Khan, S., Thainimit, S., Kumazawa, I., & Marukatat, S. (2017, May). Text detection and recognition on traffic panel in roadside imagery. In 2017 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES) (pp. 1-6). IEEE.

[31] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie. "Coco-text: Dataset and benchmark for text detection and recognition in natural images". In arXiv 1601.07140, 2016

[32] A. Shahab, F. Shafait, and A. Dengel. "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images". In ICDAR, 2011.

[33] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "Icdar 2013 robust reading competition," in ICDAR, 2013.

[34] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, et al. "ICDAR 2015 robust reading competition," ICDAR, 2015.

[35] K. Wang and S. Belongie, Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I, ch. Word Spotting in the Wild, pp. 591–604. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[36] S. Yousfi, S.-A. Berrani, C. Garcia, Alif. "A dataset for arabic embedded text recognition in tv broadcast". In: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, IEEE, 2015, pp. 1221–1225.

[37] O. Zayene, J. Hennebert, S.M. Touj, R. Ingold, N.E. Amara. "A dataset for Arabic text detection, tracking and recognition in news videos-AcTiV", ICDAR 2015.

[38] Akallouch, M., Boujemaa, K. S., Bouhoute, A., Fardousse, K., & Berrada, I. (2020). ASAYAR: A Dataset for Arabic-Latin Scene Text Localization in Highway Traffic Panels. IEEE Transactions on Intelligent Transportation Systems.

[39] Korghond, N. K., & Safabakhsh, R. (2016, May). AUT-UTP: Urban traffic panel detection and recognition dataset. In 2016 24th Iranian Conference on Electrical Engineering (ICEE) (pp. 1678-1682). IEEE.

[40] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[41] SIMONYAN, Karen et ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[42] KRIZHEVSKY, Alex, SUTSKEVER, Ilya, et HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, vol. 25, p. 1097-1105.

[43] HE, Kaiming, ZHANG, Xiangyu, REN, Shaoqing, et al. Deep residual learning for image recognition. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.

[44] LONG, Jonathan, SHELHAMER, Evan, et DARRELL, Trevor. Fully convolutional networks for semantic segmentation. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 3431-3440.

[45] J. Dai, K. He, Y. Li, S. Ren, and J. Sun. Instance-sensitive fully convolutional networks. arXiv:1603.08678, 2016.

[46] LIN, Tsung-Yi, DOLLÁR, Piotr, GIRSHICK, Ross, et al. Feature pyramid networks for object detection. In : Proceedings of the IEEE conference on

computer vision and pattern recognition. 2017. p. 2117-2125.
[47] TIAN, Zhi, HUANG, Weilin, HE, Tong, et al. Detecting text in natural image with connectionist text proposal network. In : European conference on computer vision. Springer, Cham, 2016. p. 56-72.
[48] Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks 18(5), 602–610 (2005).
[49] Kye-Hyeon Kim and Sanghoon Hong and Byungseok Roh and Yeongjae Cheon and Minje Park. PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection. ArXiv 1608.08021, 2016.

KAOUTAR SEFRIOUI BOUJEMAA is currently a PHD candidate at Sidi Mohammed Ben Abdellah Uni-versity (USMBA). She is also pursuing a research internship at the Moroccan foundation of Advanced Science and Innovation (MAScIR). Kaoutar received in 2017 and 2015, a master's degree in Big Data Analytics and Smart Systems and a bachler's degree in computer science and mathematics, from USMBA, respectively. Her research areas of interest are computer vision, intelligent transportation systems and road safety strategies developement.

MOHAMMED AKALLOUCH Mohamed. is currently a PhD student at USMBA, Fez, Morocco. He obtained a master's degree in Big Data Analytics and Smart Systems (BDSaS) and a bachelor's degree in computer science and mathematics, from USMBA, in 2018 and 2016, respectively. His research areas are based on computer vision approaches, Advanced Driving Assistance Systems (ADAS) and intelligent transportation systems.

ISMAIL BERRADA is currently an associate professor in computer science, at Mohammed VI Polytechnic University (UM6P), SCCS, Benguerir, Morocco. In 2005, he received his PHD degree in computer science from University of Bordeaux 1, France. His research mainly focuses on Artificial Intelligence's (AI) applications in multiple domains such as cognitive radio networks, radio resource management, vehicular ad hoc network, road safety, software testing and verification. Dr. Berrada had also worked for 5 years as an assistant professor in the University of La Rochelle, France, and for 10 years in USMBA, Fez, Morocco.

KHALID FARDOUSSE is currently a computer science professor at USMBA, Fez, Morocco. He obtained a computer science PHD and a master's degree in computer science and decision making, from the Faculty of Science, USMBA, Fez, Morocco, in 2010 and 2002, respectively. Dr. Fardousse' research interests are directed towards the application of computer vision, natural scene pattern recognition and the driving behavior modeling/analysis, in intelligent transportation systems.

AFAF BOUHOUTE is currently a computer science assistant professor at the faculty of science, USMBA, Fez, Morocco. She received a PHD degree in computer science, a master degree in information system, networking and multimedia, and a bachelor degree in computer science, from USMBA, Fez, Morocco, in 2018, 2012 and 2010. DR. Bouhoute had worked as a teaching assistant for 2 years at the National School of Applied Sciences (ENSA) of Fez, Morocco. She also has a 1-year Postdoctoral experience at LaBRI, Bordeaux, France. Her research interests mainly span on modeling and analysis of the driving behavior, using different techniques and algorithms with a focus on their application in intelligent transportation systems.