Arabic Text Detection on Traffic Panels in Natural Scenes

Houssem Turki
Department of Computer Engineering, University of Sfax, Tunisia
turkihoussem@gmail.com

Mohamed Elleuch
Department of Computer Engineering, University of Manouba, Tunisia
mohamed.elleuch@fss.usf.tn

Kamal Othman
Department of Electrical Engineering, Umm Al-Qura University, Saudi Arabia
kmothman@uqu.edu.sa

Monji Kherallah
Department of Physics, University of Sfax, Tunisia
monji.kherallah@fss.usf.tn

Abstract: *Identifying and acknowledging Traffic Panels (TP) and the text they display constitute significant use cases for Advanced Driver Assistance Systems (ADAS). In recent years, particularly in the context of the Arabic language, extracting textual information from TP and signs has emerged as a challenging problem in the field of computer vision. Furthermore, the significant rise in road traffic accidents within Arabic-speaking countries has resulted in substantial financial losses and loss of human lives. This is largely attributed to the limited number of diverse datasets for traffic signs and the absence of a reliable system for TP detection. Implementing warning and guidance systems for drivers on the road not only addresses this issue but also paves the way for the integration of intelligent components into future vehicles, offering decision support for transitioning to semi-automatic or fully automatic driving based on the driver's health condition. These tasks present us with two main challenges. First, it involves developing a new Arabic dataset called the Syphax Traffic Panels dataset (STP) tailored to the diverse conditions of natural scenes gathered from "Sfax," a city in Tunisia. This dataset aims to provide high-quality images of Arabic TP. Secondly, we suggest a deep learning method for detecting Arabic text on TP by evaluating the performance of the state-of-the-art algorithms in this context. In our study, we enhance the architecture of the most successful result achieved. The experiments conducted reveal promising results, affirming the significant contribution of our dataset to this research area, and even more encouraging results stemming from the enhancements made to the proposed method. The dataset we possess is accessible to the general public on IEEE DataPort https://dx.doi.org/10.21227/5zd9-pe55.*

Keywords: *Traffic panels, scene Arabic text detection, traffic textual information, Arabic scripts in the wild, deep learning.*

Received December 21, 2023; accepted June 4, 2024
https://doi.org/10.34028/iajit/21/4/3

1. Introduction

In the last ten years, Artificial Intelligence (AI) has made a significant impact on advancing cutting-edge technologies across various fields, including health, transportation, education, and robotics [21].

Advancements in self-driving car research have been substantial [40]. Progress in AI has hastened the evolution of vehicle intelligence and autonomous driving technology, ushering in a new era of transportation where vehicles can perceive their surroundings and undertake driving tasks with varying degrees of autonomy. Due to the progress in Intelligent Transport Systems (ITS), researchers have shown keen interest in autonomous navigation for vehicles. By utilizing intelligent sensors, GPS, and deep learning algorithms, these systems can be regarded as an emerging technology. Within the transportation industry, AI is extensively employed in the creation of driving assistance and autonomous driving solutions. Advanced Driving Assistance Systems (ADAS) [24], as life-saving technologies [37], are engineered to provide a range of driving assistance features by utilizing various sources of traffic data, including automatic emergency braking, driver distraction warnings, speed adjustment, and traffic sign detection [64]. In this emerging era, drivers will continue to be significant participants, while vehicles will transform into intelligent companions that offer personalized support. Visual directives provided by traffic signs play a crucial role in autonomous navigation for vehicles. Identifying text on Traffic Panels (TP) within natural scene images is a critical task in numerous real-world applications, such as autonomous driving systems. Through the utilization of autonomous driving systems, Autonomous Vehicles (AVs) are capable of accurately perceiving and detecting traffic text in real-world environments. Typically, TP are installed alongside the road using posts. The primary purpose of traffic text detection algorithms is to prevent traffic accidents. Due to their small size, complex backgrounds, partial obstructions, and variations in lighting conditions, detecting traffic text faces numerous challenges in the wild, with the added complexity of scripts that include the Arabic

language. The task can be greatly enhanced through the progress made in deep learning. Specifically, computer vision is a pivotal technology that has played a significant role in expediting the development of ADAS and enhancing their real-time capabilities. The aim of this paper is to contribute to the enhancement of AI techniques for recognizing and extracting written content from TP in natural scenes, especially when they contain Arabic text in various scripts.

1.1. Challenges of Text in the Wild

Within the realm of computer vision, identifying and understanding text within natural images present two core challenges. These challenges find application in a wide array of fields, including video analysis for broadcasting, autonomous driving technology, industrial automation, and various other domains. Both these tasks face common and complex issues associated with how text is depicted and affected by diverse environmental factors. From an AI standpoint [6], the primary obstacle to surpassing AI challenges lies in the accessibility of top-notch data. In fact, achieving AI feats approaching human-level performance has only become achievable with the introduction of suitably curated datasets. The challenges linked to recognizing text in images captured in natural scenes can be categorized into the subsequent general categories:

- Text variability: text can exhibit a wide range of colors, orientations, sizes, typefaces, and languages.
- Design of the road panel template: including elements that bear a similarity to text, such as signs, logos, and symbols.
- Scene complexity: diversity of panels in different type of roads (highway, streets, arterial, collector, local, rural, parkways, etc.).
- Effects of distortion: the influence of image distortion stemming from various contributing factors, such as the capture angle, motion blurring, inadequate camera resolution adjustments, and partial obstructions [14, 15].

Some example of general challenges in STP dataset are illustrated in Figure 1 and additional details regarding the difficulties associated with the Syphax Traffic Panels dataset (STP) are presented in section 3.

a) Text diversity.

b) Design of the road panels.

c) Scene complexity.

d) Distortion factors.

Figure 1. Samples of general categories of challenges in STP dataset.

1.2. Arabic Script Properties in Road Panels

Arabic script is a writing system used for several languages, including Arabic, Persian, Urdu, and others. It has unique properties and features that distinguish it from other scripts. Arabic script is written from right to left, unlike most Western scripts, which are written from

left to right. This directionality is consistent in both printed and handwritten Arabic text. Also, Arabic script is a cursive script, meaning that the letters are typically joined together when writing by hand. The shape of each letter can change depending on its position within a word. In addition, Arabic script is classified as an "abjad" script, which means that it primarily represents consonants. Vowels are often indicated by diacritical marks, which are not always written, especially in everyday texts. The Arabic script contains ligatures, which are combinations of two or more letters joined together with specific, aesthetically pleasing connections. Arabic script consists of 28 basic letters. These letters are divided into two categories: "Sun" letters and "Moon" letters, which affect the pronunciation of certain words. Arabic script uses diacritical marks (tashkil) to indicate short vowels and other phonetic features in the text. Common diacritics include the fatha (a short "a" sound), kasra (a short "I" sound), and damma (a short "u" sound). Most Arabic letters change shape depending on their position within a word. Each letter has an initial, medial, and final form, and some letters connect only from one side (e.g., the right) when in their initial form. The shape of some letters changes when they are connected to other letters in ligatures. Arabic script has various contextual shaping rules that ensure proper letter connections and readability. Punctuation marks in Arabic script are also written from right to left and differ from Western punctuation marks. These are some of the fundamental properties and characteristics of the Arabic script. It is a rich and visually expressive writing system with a long history and cultural significance in many parts of the world.

Arabic scripts in natural scenes present significantly greater complexity. The principal attributes of the Arabic script comprise skewed, multi-level baselines, and multi-position joining (refer to Figure 2). These characteristics originate from the writing style and the application of continuous motion techniques for every word. In Arabic script, the length of spaces conveys important information. In a single line of text, the spaces between words function to distinguish one word from another. Meanwhile, intra-word spaces, which are comparatively narrower, occur within a single word to separate sub-words. The difficulty with Arabic fonts arises from the occasional similarity in length between intra-word and inter-word spaces. Additionally, text detection systems also face challenges related to the shape of certain letters.

a) Multi-level baselines. b) Multi-position joining. c) Words as groups of letters.

d) Multi-fonts. e) Punctuation marks. f) Shape of the same letter in the end.

g) Skewness letters. h) An Arabic letter that is not in common use. i) An inverted letter.

j) Inter word and line spaces. k) Intra-word space. l) Bold format.

Figure 2. Samples of different Arabic scripts properties from STP dataset.

1.3. The Need of Text Detection on Traffic Panels

Text is an essential means of communication and carries substantial importance in our everyday lives. It can be smoothly incorporated into different documents or scenes to efficiently communicate information [27]. Identifying text can be seen as a crucial element for a diverse array of computer vision applications, encompassing tasks such as image search that involve the complexities of Arabic text [1], robotics [38], instantaneous translation [31], sports video analysis [39], automotive assistance [34], and industrial automation [36]. The progress of AI has been speeding up the improvements in vehicle intelligence and

autonomous technology for self-driving cars, ushering in a new era of transportation where vehicles can perceive and comprehend their surroundings and execute tasks with different levels of automation. In this emerging era, drivers will retain their vital role, while cars will evolve into intelligent companions that provide personalized guidance. In this context, speech controllers enable the driver to oversee specific car functions and settings as an alternative to the conventional method of manual control. Examples of such advanced voice control technologies include Apple Siri, Amazon Alexa [71], Microsoft Cortana, Bixby, and Google Assistant. Their navigation systems communicate with centralized datasets to identify the locations of traffic congestion and to plan alternative routes to bypass them. The vehicle might also have the capability to assess the driver's and traffic conditions (including facial expressions, eye movements, vocal cues, traffic signs, barriers, etc.) to activate a warning system or assume control when the driver is fatigued or experiencing stress. These advancements in the automotive sector are primarily enabled by the diverse approaches to integrate AI into the manufacturing process and the manner in which vehicles interact with their environment. Google's significant achievement in image classification in 2014 can be attributed to the introduction of a novel image object classifier known as GoogleNet, which was trained on the ImageNet dataset [55]. Subsequent to that, new Convolutional Neural Network (CNN) structures have been created, and both researchers and industries have recognized the significance of gathering and annotating high-quality data [73]. Identifying Traffic Panels (TP) and extracting their text information are crucial assignments for Advanced Driver Assistance Systems (ADAS) [8]. Their primary purpose is to enhance driver safety by preventing them from overlooking essential traffic information (such as speed limits) and reducing distractions caused by reading panels while driving. The presence of such assistance can certainly play a role in decreasing the occurrence of severe traffic accidents. Detection of text in TP can also be utilized in constructing automated visual inspection systems for signs and panels, particularly for inventory, maintenance purposes, vehicle condition, driver behavior analysis and data [48]. Traffic prediction, an integral component of automated driving, holds a vital position in managing traffic flow. The recent focus of research has been on identifying text within road panels in scene images. Despite the numerous algorithms available, most of them tend to prioritize traffic signs while overlooking other signs that contain textual information, such as guide panels. In practice, reading text on TP proves to be a challenging task due to the diverse array of markers, different types of data, and complex textual languages. Additionally, the scarcity of comprehensive naturalistic datasets, encompassing a wide range of panels and writing styles, hinders the development of more refined methods for detecting textual signs and regulatory information.

According to statistical data, Arabic is a language spoken by a substantial population, estimated to have around 422 million speakers. Furthermore, Islam, the religion connected with the Arabic language, stands as the second-largest religion worldwide, with its adherents constituting approximately a quarter of the global population. From an analytical perspective, it can be deduced that a significant proportion of Muslims possess an understanding of the Arabic language, largely due to the fact that the Holy Quran is written in Arabic. Arabic is both the native language and the official language in many nations. In recent times, there has been a noticeable rise in the number of Arabic speakers. Nonetheless, despite this expansion, the field has seen limited research and development due to specific complexities and challenges linked to the Arabic language. Building a dependable Recognition System (RS) for cursive languages, such as Arabic, poses a significant challenge. Moreover, the severe scarcity of a standardized, comprehensive, and openly accessible Arabic dataset within the domain presents a considerable obstacle for research on Arabic text detection and recognition. The challenge lies in the complexities of text detection and dataset annotation.

Our primary goal is to tackle the shortage of available datasets containing text on traffic panel images taken in the wild conditions and featuring Arabic scripts, while also improving the most effective approach for Arabic text detection in this context. The STP dataset is presented as a remedy for this issue, providing a realistic compilation of images captured in "Sfax," a city located in southern Tunisia and the capital of the region [66]. The STP dataset includes a wide array of textual information found in images featuring traffic signs in natural environments. These images encompass various types of signs, including traditional road signs, electronic displays, signs made from different materials, panel lights, banners, and more, as depicted in Figure 3. These images were collected under different weather conditions and at various times of the day, as shown in Figure 6. Furthermore, it incorporates images from the wild containing text that were collected manually, featuring a wide range of Arabic text variations such as different fonts, colors, sizes, and various levels of intricate noise. This introduces new challenges that are not present in other datasets in this domain. The potential result of our novel dataset and the improved method would be particularly beneficial for tourists traveling to Arabic-speaking countries who are not acquainted with the Arabic language. Moreover, it holds the promise of enhancing road safety on roadways in Arab nations and facilitating the integration of intelligent components in future vehicles, providing decision support and the capability to transition to semi or fully automated driving depending on the driver's health status. Furthermore, there are road panels and

notifications positioned alongside the road to provide directions and alert drivers to possible dangers or risks they might face while driving. To the best of our knowledge, this dataset, which concentrates on Arabic scripts and features natural scene images with manual text annotations in the Arabic language, is the first publicly available collection of its kind. It also encompasses all the prevalent challenges encountered in datasets in this particular domain. This research suggests enhancements similar to the various datasets designed for Robust Reading competitions and datasets that consist of images taken from natural scenes.

This work puts forth four contributions:

1. Unlike numerous specialized datasets found in existing literature, the present study introduces a new dataset focused on TP in the wild. This dataset is tailored to Tunisian templates and road terminology, encompassing Arabic inscriptions found on various roads. The STP dataset is openly accessible and comprises high-quality images captured individually in natural scenes.
2. Due to the specific Arab environment in North African countries and varied climate change, we include different panels and Km-points challenges in images of natural scenes containing the Arabic scripts.
3. To test state of the art text detection deep learning techniques, we present comparative results of 3 deep methods a multi-purposes dataset applied to our new dataset and other datasets in the field that covers Text in the Arabic traffic panel.
4. We enhance the structure of the best outcome attained through the utilization of the comparative methods.

a) Printed text. b) Printed text. c) Printed text. d) Printed text.

e) Sticker text. f) Sticker text. g) Electronic text. h) 3D text.

i) 3D text. j) Hand-written. k) Banner. l) Symbols.

Figure 3. Samples of diverse Arabic text information patterns in the wild from STP dataset.

The rest of the paper is structured as follows: section 2 provides a summary of the present datasets used for text scene detection, which includes datasets for TP as well as existing Arabic datasets in the field. Section 3 contains extensive details and data about the STP dataset, covering its description and relevant statistics. In section 4, a concise overview is given regarding the evaluation methods and metrics utilized in the study. Section 5 is focused on the discussion of the achieved results, and the final remarks are outlined in section 6.

2. Related Works

Deep learning has garnered widespread attention due to its capability for learning from data. Given this, methods based on Convolutional Neural Networks (CNNs) have been instrumental in the detection of traffic panel features. Zhu *et al*. [76] utilized a Fully Convolutional Network (FCN) to detect traffic signs and employed Deep Convolutional Neural Networks (DCNNs) for their classification. Meng *et al*. [32] in their work identified traffic signs using the Single Shot multibox Detector (SSD) approach with the utilization of image pyramids. Li *et al*. [23] enhanced the effectiveness of small object detection by leveraging Generative Adversarial Networks (GANs). All these approaches were capable of detecting certain traffic signs. Nonetheless, due to the diminutive size of traffic signs and the constraints of available datasets, it proved challenging for these methods to strike a well-balanced compromise between accuracy, comprehensiveness, and real-time processing. In certain research, there has been an emphasis on employing template-based data augmentation to attain an extensive detection of traffic signs. Bloice *et al*. [3] introduced a pipeline-based method for image augmentation, which encompassed

elements such as randomized elastic distortions and z-stack augmentation. These methods are accessible to the public. In [77] a novel method introduced, involving the training of two separate CNNs for the dual tasks of traffic sign detection and simultaneous traffic sign detection and classification. Li *et al*. [22] recommended a framework that combines Faster R-CNN and MobileNet with refinements to accurately locate all traffic signs. Zhang *et al*. [72] introduced a cascaded Region-based Convolutional Network method (R-CNN) approach with multi-scale attention, designed to address the issue of similar traffic signs. Wu *et al*. [67] introduced a two-level detection model, composed of a location phase and a classification phase, both of which implemented the YOLOv3 [41] architecture. Sermanet and LeCun [46] introduced a multi-scale feature CNN for the purpose of classifying traffic signs, employing layer skipping connections. Jurisic *et al*. [17] put forward a CNN network model designed to classify multiple datasets. Haque *et al*. [13] introduced an innovative, energy-efficient CNN architecture for traffic sign recognition. This architecture comprises four layers with a limited number of feature maps per layer and incorporates only one hidden fully connected layer. Turki *et al*. [60] we return to these algorithms in detail in section 4 of this paper. These approaches have also been employed on various datasets for text detection in natural scenes, which include a limited number of road panel images and predominantly concentrate on Latin scripts. Widely used Latin datasets such as COCO-text [61], ICDAR [18, 19, 47], and Street View Text database (SVT) [65] are examples of this. These datasets primarily consist of English text gathered from diverse sources like announcements, books, and posters. In the case of Arabic scripts, there are only a limited number of benchmark datasets, such as ALIF [68] and AcTiv [69], which are assembled from Arabic news channels, as well as ARASTI [58] and ICDAR2017-MLT [35], a comprehensive multilingual text collection project encompassing nine different languages. In our work we focus only on the datasets of road panels and especially those that contain text in the Arabic language.

2.1. Traffic Signs Datasets

In order to facilitate the advancement of robust techniques for traffic sign recognition, various benchmark datasets have been put forward. The German Traffic Sign Recognition Benchmark (GTSRB) [53] stands as one of the most extensive and diverse datasets for traffic sign recognition. The dataset was gathered from roadways in Germany and comprises over 50,000 images of traffic signs. LISA [33] is a compilation of video clips and annotated images depicting U.S. traffic signs, featuring a total of 7,855 images representing 47 distinct traffic sign types. Tsinghua-Tencent 100k [77] is a Chinese dataset focused on traffic sign detection. It stands out as one of the most extensive datasets to date, offering 100,000 images encompassing 30,000 instances of traffic signs. The Russian Traffic Sign Dataset (RTSD) is another notably large collection, comprising 179,138 labeled frames featuring 156 distinct sign categories [7]. The Swedish traffic sign detection and recognition database consist of 20,000 frames captured across 350 kilometers of Swedish highways and urban roads [20]. The Belgium traffic sign database bears a resemblance to the German datasets in its content and structure [57].

These datasets were primarily designed for the purpose of traffic sign detection and recognition, and as such, they may not be suitable for extracting text from TP, particularly guide panels. In response to the absence of available datasets for traffic guide panels, Rong *et al*. [44] gathered a new challenging dataset that focuses on guide panels located on U.S. highways. Gonzalez *et al*. [12] utilized Google street view to generate a dataset comprising two distinct Spanish highways, which served as the basis for validating their extraction technique.

2.2. Arabic Traffic Panels Datasets

In the field of literature, the frequently recommended method for identifying TP relies on the utilization of color and shape segmentation techniques [63]. Unfortunately these latter methods fail to satisfy the efficiency criteria for real-time applications. As a result, the proposed approach for extracting text from TP involves training a text detector/recognizer using an outdoor scene text dataset. There are a limited number of these collections, and they are primarily gathered in a road-related context. Our research is specifically centered on Arabic text-based traffic panel datasets, such as the ASAYAR dataset [2], which comprises 1,763 images gathered from various Moroccan highways. These images were extracted from videos of two complete journeys and meticulously annotated with 16 different object categories. The ASAYAR_TXT dataset comprises 1,375 images containing text-based panels, and it includes annotations at both the word and line levels, formatted according to Pascal VOC [10]. The second dataset designed for the detection of Arabic text-based TP is the ATTICA dataset [4], which consists of 1,215 images. This dataset includes 3,173 bounding boxes for TP, 870 for traffic signs, and 7,293 for Arabic text. The images were sourced from open-access internet images.

It is important to highlight that the STP dataset's significant contributions and difficulties include the incorporation of authentic Arabic text scripts, which encompass many of the challenges found in state-of-the-art datasets, as outlined in Table 1.

Table 1. Various challenges of Arabic scripts in STP dataset.

Challenges	Description
Scene complexity	Can be highly complex and diverse, can include cluttered backgrounds, varying lighting conditions, perspective distortion, occlusions, and uneven text sizes.
Text size variability	Wide range of sizes, from very small to very large.
Low text resolution	Some images.
Crowded text and overlapping instances	Contain multiple text instances in close proximity, overlapping with each other or with other objects.
Blurry or unfocused text	Some images.
Multi-lingual text	Multi-lingual and multi-script text.
Manual annotation	Multi-classes.
Real-world variation	Lighting conditions.
Complexity of text instances	Varying sizes, fonts, curved, tilted, or heavily occluded.
Varied backgrounds	Clutter, texture, and varying lighting conditions.
Size and aspect ratio	Some text may be very small.
Irregular Arabic text shapes	Bent text.
Low contrast	Some images.

3. Dataset Construction

In the following section, we will delve into the process of collecting and annotating the data, concluding with crucial statistical information about the STP dataset. For more in-depth information, consult reference [59]. It's noteworthy that the dataset introduced in this research is available to the public. Table 2 offers a comparison between the STP dataset and other Arabic traffic panel datasets that incorporate Arabic scripts, emphasizing the strengths and benefits of the STP dataset.

Table 2. Text detection in Arabic TP datasets.

Detection dataset	ASAYAR_TXT [2]	ATTICA_TEXT [4]	STP [59]
Size	1375	1180	506
NB of classes	7	4	3
Script	Arabic, Latin	Arabic, Latin	Arabic, Latin
Categories	Line level, word level	Line level, Word level	Line level
Availability	Public	Public	Public
Source device	Camera, Mobile phone	Internet	Camera
Source data	Video capture	Download images	Photography
image-quality	High-quality	Medium and bad-quality (unreadable line)	High-quality
Km-Point (see Figure 5)	No	Yes	Yes
Source of roads	Highway	Highway, streets, local	Highway, streets, arterial, collector, local, rural, parkways, etc.
Add-Panel(on sides of highways, joined to TP and traffic signs), (see Figure 5)	No	Yes	Yes
Other pattern panel, (see Figure 3)	No	Yes	Yes
Panel patterns.	Unique	Various	Various

3.1. Data Collection

The STP dataset was collected in "Sfax," the second-largest city in Tunisia, after the capital city. A total of 506 images were obtained through manual collection, with each image representing challenges in text detection that reflect the real complexities encountered in 15 distinct routes (Tunis main road, Sidi Mansour, Mahdia road, downtown, Sakiet Eddayer, Nasryia, Taniour, Sakiet Ezziet, Elayn, Gremda, Chehia, Manzel Chaker, Lafran, Gabes road, Matar road) in addition to ring roads, roundabouts, intersections, airport and highways. The total approximate distance covered was around 470 kilometers. All the routes mentioned, which were traversed, included TP on both the outbound and return journeys (refer to Figure 4). The TP feature Arabic text, providing directional information and indicating the distance in kilometers to the designated destination. These panels are frequently encountered at intersections, divergent lanes, and across the city. The different road signs in Tunisia follow a design template according to fixed standards from the Ministry of Transport in all cities of Tunisia. Our primary goal is to identify Arabic text in various types of panels in the wild (refer to Figure 5).

We employed a Reflex Canon EOS 2000D camera with a 24.1-megapixel sensor and two stabilized lenses: the EF-S 18-55 mm f/3.5-5.6 IS II and the EF 75-300 mm f/4-5.6 III, for capturing the images. In specific instances, a tripod was used, depending on the shooting location. The image dimensions and resolutions range from 1200x788 to 4000x3000 pixels. Both horizontal and vertical resolutions were consistently set at 72 dots per inch (dpi).

a) Highway. b) Highway. c) Crossroads. d) Crossroads.

Figure 4. Examples of different roads of Sfax city traced and included in the STP dataset.

a) Fixed panel. b) Wall panel. c) Km-point.

d) Add-Panels. e) Add-Panel. f) Add-Panel.

g) Add-Panel. h) Removable panels. i) Mixed-panel text and sign.

Figure 5. Examples of different type of panels from STP dataset.

Detecting objects in an outdoor setting can be affected by various environmental factors, including variations in daylight conditions at different times of the day, such as dawn, daytime, dusk, or nighttime. The camera was used for individual shots while on foot, while driving a car, or while accompanying a car driver. Certainly, engaging in photography while driving in heavy traffic can pose significant risks. The positioning of the camera angles and making resolution adjustments added extra difficulties, further enhancing the development of a robust dataset. Additionally, during the collection process, we took into account other factors, such as adverse weather conditions, such as rain and intense sunlight, as depicted in Figure 6.

a) Sunny weather. b) Normal weather. c) Cloudy weather. d) Rainy weather.

e) Early Morning. f) Daytime. g) Evening. h) The night.

Figure 6. Samples of TP pictures captured during different times of the day and in varying weather conditions.

3.2. Data Annotation

Three researchers dedicated a period of five months to manually annotate the STP dataset, involving a collective commitment exceeding 800 hours. The Labelme tool [62] was employed for annotating the images. The tool referenced in [10] automatically produces an XML metadata file in Pascal VOC format for every image within the STP dataset. The generated files are required to share identical names with their corresponding images. The rationale behind choosing these two annotation formats lies in their widespread usage among researchers involved in developing advanced detection algorithms. The XML metadata file, formatted in Pascal VOC, comprises the image name, class designation, and the bounding box coordinates.

In order to cater to a variety of potential applications for the dataset, the annotation process encompassed multiple class categories. The STP dataset comprises a total of 506 images, with 106 allocated for testing and 400 for training, representing 20% and 80% of the total images, respectively. On average, each image contains approximately 2.67 bounding boxes. The images consist of text in Arabic scripts, and only the Arabic script is annotated. To construct a dataset conducive to line-level text detection, five label classes were employed:

a) Arabic. b) Number. c) Mixed.

Figure 7. Line-level classes of STP dataset.

For line-level annotation, three categories were utilized: Arabic, Number, and Mixed, where “Mixed” refers to text that incorporates both the Arabic language

and numbers within the same line (as illustrated in Figure 7). Table 3 displays the count of boxes for each class, while Figure 8 offers a graphical representation of the statistics. From these statistics, we can deduce that:

Table 3. The count of bounding boxes for each class of STP dataset.

Line Level				
Class	**Number of boxes**			**Average by image**
	Test	**Train**	**Total**	
Arabic	216	766	982	1.94
Number	34	131	165	0.32
Mixed	42	162	204	0.40
	Total of boxes		**1351**	

1) A cumulative total of 1351 bounding box objects were annotated.
2) Both Arabic and Number exhibit relatively uniform distribution and fairly evenly at the line levels.

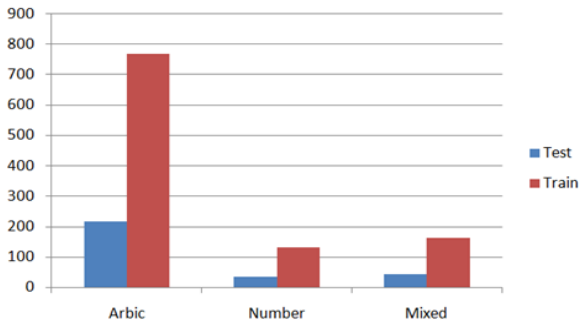

Figure 8. Number of boxes per class in STP dataset.

3.3. Challenges

Engaging with the STP dataset for detecting natural scene text poses several notable challenges. Indeed, accurately recognizing Arabic text poses a formidable challenge, largely attributed to the imperative of precisely framing the image during the capturing phase. To attain enhanced stabilization with either a stationary or moving camera, it is essential to choose a comfortable position and consistently fine-tune camera settings such as size and quality, picture style, frame rate, shutter speed, aperture, white balance, and ISO. These modifications become essential since automatic camera capture might not always be effective, especially considering the existing conditions during the shooting process. These modifications are also justified to concentrate on capturing the target area while reducing potential obstacles that could impede the process.

Three classes of obstacles are present (refer to Figure 9):

- Objects: such as trucks, electrical panels, TP, etc.
- Natural obstacles: light source, trees, sunny weather, etc.
- Shooting conditions: shooting angle, incomplete framing, capture frame area, car windshield internal reflection, traffic and dangerous paths, etc.

a) Objects of obstacles.

b) Natural obstacles.

c) Shooting conditions.

Figure 9. Samples of images in STP dataset containing problems during the shooting.

Dealing with the STP dataset is a considerable challenge due to the diverse and intricate features of text images present in natural scenes. These difficulties can be categorized into five primary aspects, as illustrated and elucidated in Figure 10:

- Panels in poor conditions: soiled, old or improperly installed.
- Text orientation: horizontal, curved, multi-oriented, skewed or partially occluded.
- The extensive variety and variability of text within panels were carefully addressed during dataset construction to ensure effective handling of diverse sizes, font styles, electronic text panels and colors. Additionally, the text in the image may be positioned either in close proximity to or at a distance from the camera.
- The intricate background of panels: managing

backgrounds in natural scenes is frequently challenging or impractical, noise, motion blurs, featuring intricate designs and various other types of interference.

- The image capture environment is less than ideal or imperfect: obstacles in the captured image may cause damage to text instances, thereby reducing text recognition accuracy.

BP, CT, ST, CB, MS, DC
DT, M, FoP, FP, M, IV, MS
MB, SmT, MC, ML, MF, ST, MS, CC, SmT
OT, ML, OP, PO, PO, ST, SmT, CC

Figure 10. Samples of images in STP dataset containing various challenges in the wild. BP: Broken Panel, CT: Curved Text, ST: Skewed Text, CB: Complex Background, MS: Multi-Sizes, DC: Dirty Characters, DT: Dirty Text, ML: Multilingual text, FoP: Folded Panel, FP: Fallen Panel, IV: Illumination, Variation, MB: Motion Blur, SmT: Smal Text, MC: Multi-Colors, MF: Multi-Fonts, CC: Close Characters, OT: Old Text, OP: Old Panel, PO: Partial Occlusion.

4. Algorithms of Text Detection on Traffic Panels

To assess the practicality and efficacy of the STP dataset, we exclusively employ a baseline comprising various state-of-the-art neural network architectures for text detection, namely Single-Shot Oriented Scene Text Detector (TextBoxes++), Connectionist Text Proposal Network (CTPN), and Efficient and Accurate Scene Text Detector (EAST). While there are other algorithms utilized in object detection and recognition within the field, our study specifically concentrates on text detection in road panels.

4.1. Text Detection Methods

In images depicting natural scenes, two types of techniques for detecting text can be observed:

The initial group of approaches relies on the concept of region proposals. The method utilizes a conventional target detection network as its fundamental model but enhances its functionality by incorporating text detection for practical applications in the real world. The method transforms the original multi-class target detection model into a consolidated single-class detection model specifically designed for recognizing instances of text. Various deep neural network algorithms proficiently extract image features and yield satisfactory outcomes in detection, including YOLO [42], RCNN [11], Faster-R CNN [43], SSD [29], and similar methods. Moreover, there are notable approaches such as CTPN [56], TextBoxes++ [26], ABCNet [49], SegLink [52], etc., which fall within this class of methods.

The second group of approaches relies on semantic segmentation. The primary emphasis of the text detection algorithms within this classification revolves around semantic segmentation. Advancements in these algorithms have been achieved through enhancements to the FCN [30] and FPN [28] methodologies. The method extensively utilizes deep convolution and up sampling techniques to accomplish multi-level fusion. Through the application of these techniques, the approach can determine if each pixel in an image corresponds to a text region and obtains pixel-level label predictions. Various segmentation networks commonly employed for text detection include EAST [75], PixelLink [9], PSENet [25], and other similar methods.

4.2. Algorithms Used with Arabic Scripts

Based on statistical data, it has been determined that approximately 422 million people use the Arabic language for communication. Moreover, Islam is acknowledged as the second-largest global religion, with approximately a quarter of the world's population identifying as followers of this faith. Nevertheless, there is a lack of research utilizing algorithms for the identification of Arabic text in the wild, primarily due to specific complexities and challenges inherent in the scripts. For several years, researchers have been scrutinizing Arabic records in various forms, such as printed, handwritten, and Optical Character Recognition (OCR). Historically, a considerable focus among researchers has been placed on the segmentation of words within Arabic scripts. However, researchers are currently focused on the identification and segmentation of free text in Arabic [16].

Numerous systems for recognizing text in natural scene images have been developed. Yousfi *et al*. [68] introduced a method that identifies individual characters in English scripts, followed by the application of a Deep Convolutional Neural Network (DCNN) model for recognition. Butt *et al*. [5] have presented a CNN-Recurrent Neural Network (RNN) model incorporating attention mechanisms to detect Arabic text in images depicting natural scenes. Jain *et al*. [16] proposed a unified CNN-RNN model designed for the recognition of Arabic text in both natural scenes and videos. Turki *et al*. [60] have introduced an approach for text detection that depends on Maximally Stable Extremal Regions (MSER) and features from CNN. Akallouch *et al*. [2] utilize TextBoxes++ [26], CTPN [56], and EAST [75] techniques for the ASAYAR and ATTICA datasets [2, 4].

Ongoing research within the domain of pattern recognition persists in the quest for advancements in Arabic text recognition.

4.3. Methods Used in our Experimental Study

To confirm the viability and efficiency of the STP dataset, we utilize various state-of-the-art neural network architectures for text detection. Among our foundational models are CTPN [65], EAST [75], and TextBoxes++ [26]. The reason for choosing these three methods lies in their promising results within the realm of text detection on traffic signs, especially considering the enhancements introduced in recent and improved versions [50, 54, 74].

To overcome the size limitation problem in our STP dataset in the field of text detection in natural scenes can be achievable with approaches such as transfer learning and data augmentation.

The three architectures of CTPN, EAST and TextBoxes++ chosen are models already pre-trained on other benchmarks. Therefore, fine-tuning these models on our specific dataset can significantly improve performance even with limited data.
Additionally, data augmentation techniques augment our existing dataset to create variations and can help diversify data.

a) CTPN

Tian *et al*. [56] introduced a method known as the Connectionist Text Proposal Network (CTPN), which interprets text regions as sequences of connections formed by various elements. Following this, they applied Recurrent Neural Networks (RNNs) to extract features encoded sequentially, which are then used for regression predictions. The architecture of the network model is configured to recognize textual content across different scales and languages. CTPN employs VGG16 for extracting features and incorporates the anchor regression technique derived from Faster-RCNN [43]. This enables the Region Proposal Network (RPN) to identify objects of varying dimensions using a uniform sliding window size, offering flexibility in parameter adjustments. By employing deep convolutional networks and parameter sharing, CTPN can efficiently locate lines of text. However, it faces limitations in recognizing text that is not horizontally aligned; the majority of text images in the STP dataset are oriented horizontally, reflecting a common characteristic of naturally occurring text images.

b) EAST

Zhou *et al*. [75] introduced the Efficient and Accurate Scene Text Detector (EAST) algorithm as an effective and precise approach for text detection within a scene. The algorithm employs a single neural network and has the ability to predict lines or words of text in any orientation within an image. This eliminates the need for the challenging task of compiling candidates and segmenting words among them. The EAST approach utilizes the FCN [30] model to directly produce text regions. The acquired images were then subjected to the Non-Maximum Suppression (NMS) stage to produce the final results. The architecture of the EAST algorithm is uncomplicated, leading to enhancements in both speed and accuracy.

c) TextBoxes++

The Single-Shot Oriented Scene Text Detector (TextBoxes++) is an end-to-end neural network specifically created for the detection of text in scene images. It draws inspiration from the Single Shot multibox Detector (SSD) network and is notably swift and efficient in terms of training duration. The architecture of TextBoxes++ comprises a FCN consisting of 13 layers from the VGG16 base model, followed by an additional 10 convolutional layers. Additionally, there are 6 text-box layers connected to 6 intermediate convolution layers of the base model. At every position in a feature map (derived from an

intermediate convolutional layer), the text-box layer divides K default boxes and produces a 2-dimensional score vector indicating the presence or absence of text, along with an N-dimensional coordinate vector for each box. The value of N can be 4 for rectangle bounding boxes, 5 for rotated boxes, or 8 for quadrilateral boxes.

5. Evaluation Metrics

To assess the applicability of the STP dataset, we have implemented a method consisting of the following stages:

- Dividing the sub-datasets into separate training and testing sets involved utilizing the technique of stratified sampling. The datasets were split into training (80%) and testing (20%) subsets. The training dataset includes 400 samples, while the testing dataset comprises 106 samples. This methodology was applied to both word-level and line-level datasets.
- Selection of a specific data augmentation technique.
- Choosing the most advanced models for utilization with the STP dataset.
- Establishing appropriate evaluation metrics for each selected model.
- Performing a benchmarking process on the chosen algorithms and thoroughly analyzing the obtained results.
- Improving the methodology that yielded the best results.

5.1. Choose of a Specific Data Augmentation Techniques

Employing effective data augmentation methods is beneficial when training models for the detection of text on TP. These methods increase the diversity and quantity of training data, thereby improving the models' ability to generalize in the wild. We apply these techniques to the STP dataset [59]. Selecting an appropriate augmentation technique is crucial, as it should closely resemble the characteristics of genuine images found in natural scenes and align with the attributes of the chosen dataset. By applying text detection methods in this manner, we can achieve positive results. We have selected four data augmentation techniques that cover fundamental image modifications [51]. To start, we opted to employ geometric transformations through rigid transformations [45] to generate two extra skew angles that closely replicate challenging camera shooting perspectives; one to the right and one to the left. Secondly, we incorporated a horizontal directional blur effect to simulate the capture of moving images from a camera. Thirdly, we applied color space transformations by adjusting white balance. This approach enabled us to generate two distinct variations in color temperature and brightness, representing different times of the day. Lastly, we introduced noise injection from a Gaussian distribution, coupled with a 20% increase in contrast, histogram equalization, and sharpening.

5.2. Metrics

In this section, we assess the appropriateness of the proposed dataset by testing three approaches for detecting text in natural scene images, CTPN [56], EAST [75], and TextBoxes++ [26]. Table 4 describes the functionalities of each model used (Backbone and backend) and the level of text detection (line and word).

Table 4. Backbone and backend employed in every model.

Model	Backbone	Backend	Used data
CTPN	VGG16	TensorFlow	Z
EAST	VGG16	TensorFlow	Line
TextBoxes++	VGG16	TensorFlow	Line

The evaluation metrics are the precision, recall and F-score defined as:

$$Precision = \frac{\sum_i^N \sum_j^{|D^i|} M_D(D_j^i, G^i)}{\sum_i^N |D^i|} \quad (1)$$

$$Recall = \frac{\sum_i^N \sum_j^{|G^i|} M_G(G_j^i, D^i)}{\sum_i^N |G^i|} \quad (2)$$

$$F\text{-}score = 2 \, X \, \frac{Precision \; X \; Recall}{Precision + Recall} \quad (3)$$

N represents the complete count of images within a given dataset. $|D^i|$ and $|G^i|$ are the number of detection and ground true rectangles in i^{th} image. $M_D(D_j^i, G^i)$ and $M_G(G_j^i, D^i)$ are the matching scores for detection rectangles D^j and ground true rectangle G^j.

To enhance performance, we utilized the TensorFlow framework to create diverse models. The pre-trained models (backbones) utilized in the experiments are publicly available and can be accessed on Github. We conducted training and evaluation of the diverse models on a personal computer, specifically the MSI THIN GF63, Intel Core i7-12650H, 24 Mo of cache memory, 4,70 GHz Turbo max, 10-Cores, 512 Go SSD, RAM Memory: 16 Go DDR4-3200, Graphic card: NVIDIA GeForce RTX 4050.

5.3. Analysis and Discussion

The effectiveness of the three datasets was assessed by applying metrics like Precision, Recall, and F-score. During our training process, we utilized level data for CTPN, EAST and TextBoxes++. Table 5 illustrates that CTPN achieved the highest recall and F-score with the three datasets. Table 6 and Table 7 illustrates that EAST achieved the highest Precision and F-score with the three datasets. Table 5 of STP dataset achieved the second best result with EAST method and the first best result with CTPN method (a precision of 55%, recall of 73% and F-score of 63%). The number of road panels images in STP dataset is limited but it is improved with the specific data augmentation techniques applied, in addition the quality of the images is high and despite

there being a variation in the sizes of the text and the multiple existing challenges already mentioned previously, the concentration of capture on the Arabic text as a single language made it possible to reduce the drop in the result.

Table 5. Experimental result on the STP Dataset.

Method	Precision (%)	Recall (%)	F-score (%)
CTPN	0.55	**0.73**	**0.63**
EAST	**0.59**	0.47	0.52
TextBoxes++	0.28	0.39	0.33

Therefore, the use of specific augmentation techniques applied made it possible to fill the gaps in the creation of the database such as the variation of lights and shooting conditions. Applying data augmentation techniques for text detection in images of natural scenes contributes to several considerations: text variability, spatial transformations, noise and distortions, and background variations.

Table 6 of ASAYAR_TXT dataset achieved the highest result with EAST method (a precision of 71%, recall of 89% and F-score of 79%). The images of road panels in ASAYAR_TXT are homogeneous (unique template) and in high quality, in addition the size of the texts is readable with less challenge than the other two datasets. Finally, Table 7 of ATTICA_TXT dataset achieved the second best result with CTPN method and the first best result with EAST method (a precision of 48%, recall of 66% and F-score of 56%). The images of road panels in ATTICA_TXT dataset are heterogeneous and the majority of images downloaded are of medium or poor quality (the text is not readable). Overall, the results obtained with CTPN and EAST methods applied on ASAYAR_TXT and STP datasets are close. We try to improve these results on the three datasets in parallel using an improved version of CTPN method because if we choose between CTPN and EAST, our major contribution focuses more on Arabic scripts as a single language annotated in STP dataset containing various challenges in the wild (ASAYAR dataset is in multilingual). The results related to text detection on the three datasets are illustrated in Figure 11. Figure 12 displays the examples of text detection failures. It is evident that the primary factor leading to the failure of text detection on TP is the fluctuation in light reflection, which is contingent on the time of day and the conditions of the panels.

Table 6. Experimental result on the ASAYAR_TXT.

Method	Precision (%)	Recall (%)	F-score (%)
CTPN	0.67	0.85	0.75
EAST	**0.71**	**0.89**	**0.79**
TextBoxes++	0.44	0.63	0.52

Table 7. Experimental result on the ATTICA_TEXT.

Method	Precision (%)	Recall (%)	F-score (%)
CTPN	0.41	**0.69**	0.51
EAST	**0.48**	0.66	**0.56**
TextBoxes++	0.22	0.44	0.29

a) Text in add-panel. b) Text in highway. c) Text in sign.

Figure 11. STP, ASAYAR and ATTICA dataset's test results with CTPN model respectively in image (a), (b), and (c).

a) Distant text. b) Illuminations. c) Low light.

Figure 12. Failed text detection samples.

5.4. Improvement of the Best Results Method Obtained (CTPN)

Our enhancements involve refining the side-refinement detection frame merging mechanism, incorporating height information into the detection location and merging process, and substituting the BiLSTM network with GRU [70]. This accelerates both network training and application runtime, resulting in improved network efficiency. The enhancement in the side-refinement merge mechanism demands special attention, particularly in tasks such as Detecting Text in Fine-scale proposals and recurrent connectionist text proposals. The purpose of the side-refinement stage is to consolidate and summarize the located "small rectangles" to acquire the positional information of the necessary text information tailored to Arabic scripts. In the enhanced CTPN method, the recurrent connectionist text proposals stage employs a bidirectional LSTM pair

to extract features from both the VGG16 feature extraction layer and the feature obtained from the 3*3 sliding window. Subsequently, it conducts feature sequence prediction.

The novel approach significantly enhances the algorithm's efficiency. We implement the improved algorithm on the STP, ASAYAR_TXT, and ATTICA_TXT datasets, yielding promising results (refer to Table 8).

Table 8. Experimental result using an enhanced CTPN.

Dataset	Precision (%)	Recall (%)	F score (%)
STP	0.60	0.81	0.69
ASAYAR_TXT	**0.75**	**0.88**	**0.81**
ATTICA_TXT	0.52	0.76	0.62

6. Conclusions

This work presents the STP dataset, a newly released dataset designed exclusively for identifying Arabic text in images captured from natural scenes, particularly on TP. The dataset is openly available for researchers in the community to use [59]. The STP dataset consists of a collection of high-quality images meticulously taken in the natural environment of the Tunisian city "Sfax". These images cover diverse categories of Tunisian road panels, incorporating Arabic scripts. We have provided a comprehensive description of the approach used for collecting and annotating the data. Our results suggest that STP functions as a reliable primary data source for detecting text in scenes TP. Modern text detection techniques evaluated on the dataset have demonstrated favorable results. To our knowledge, this dataset is the initial of its kind publicly available, encompassing the most challenging aspects of text detection in the wild, with a specific emphasis on Arabic scripts within natural scenes. It has been systematically assembled, incorporating a variety of scripts, different levels of difficulty, and thorough annotations in the Arabic language. The dataset incorporates Arabic scripts and encompasses all the challenges present in other comparable datasets within the same domain. Our upcoming plan involves enlarging the dataset by adding samples from more Tunisian cities. In addition, a considerable extension will soon be carried out in future cooperative work with other research teams in Arab countries with a particular focus on enhancing the recognition of Arabic script for text identification on TP.

Acknowledgement

We would like to express our sincere gratitude to Sfax University, Sfax Governorate, and the Tunisian Ministry of Interior for their invaluable support in this research. Their issuance of photography licenses for public spaces, conducted in accordance with personal data regulations, significantly eased the data collection process.

References

[1] Ahmed S., Razzak M., and Yusof R., *Cursive Script Text Recognition in Natural Scene Images*, Springer, 2020. https://doi.org/10.1007/978-981-15-1297-1_2

[2] Akallouch M., Boujemaa K., Bouhoute A., Fardousse K., and Berrada I., "ASAYAR: A Dataset for Arabic-Latin Scene Text Localization in Highway Traffic Panels," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3026-3036, 2022. DOI:10.1109/TITS.2020.3029451

[3] Bloice M., Roth P., and Holzinger A., "Biomedical Image Augmentation Using Augmentor," *Bioinformatics*, vol. 35, no. 21, pp. 4522-4524, 2019. https://doi.org/10.1093/bioinformatics/btz259

[4] Boujemaa K., Akallouch M., Berrada I., Fardousse K., and Bouhoute A., "ATTICA: A Dataset for Arabic Text-Based Traffic Panels Detection," *IEEE Access*, vol. 9, pp. 93937-93947. 2021. DOI:10.1109/ACCESS.2021.3092821

[5] Butt H., Raza M., Ramzan M., Ali M., and Haris M., "Attention-based CNN-RNN Arabic Text Recognition from Natural Scene Images," *Forecasting*, vol. 3, no. 3, pp. 520-540, 2021. https://doi.org/10.3390/forecast3030033

[6] Chen C., Wang C., Liu B., He C., Cong L., and Wan S., "Edge Intelligence Empowered Vehicle Detection and Image Segmentation for Autonomous Vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 13023-13034, 2023. DOI:10.1109/TITS.2022.3232153

[7] Chigorin A. and Konushin A., "A System for Large-Scale Automatic Traffic Sign Recognition and Mapping," *in Proceedings of the CMRT13-City Models, Roads and Traffic*, Antalya, pp. 13-17, 2013. https://doi.org/10.5194/isprsannals-II-3-W3-13-2013

[8] Cleofas-Sánchez L., Posadas-Durán J., Martínez-Ortiz P., Loyo-Desiderio G., Ruvalcaba-Hernández E., and González Brito O., "Automatic Detection of Vehicular Traffic Elements Based on Deep Learning for Advanced Driving Assistance Systems," *Computación y Sistemas*, vol. 27, no. 3, pp. 643-651, 2023. https://doi.org/10.13053/cys-27-3-4508

[9] Deng D., Liu H., Li X., and Cai D., "Pixellink: Detecting Scene Text Via Instance Segmentation," *in Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, pp. 6773-6780, 2018. https://doi.org/10.1609/aaai.v32i1.12269

[10] Everingham M., Gool L., Williams C., Winn J., and Zisserman A., "The Pascal Visual Object Classes Challenge," *International Journal of*

Computer Vision, vol. 88, no. 2, pp. 303-338, 2010. https://doi.org/10.1007/s11263-009-0275-4

[11] Girshick R., Donahue J., Darrell T., and Malik J., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, pp. 580-587, 2014. DOI:10.1109/CVPR.2014.81

[12] Gonzalez A., Bergasa L., and Yebes J., "Text Detection and Recognition on Traffic Panels from Street-Level Imagery Using Visual Appearance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 1, pp. 228-238, 2014. DOI:10.1109/TITS.2013.2277662

[13] Haque W., Arefin S., Shihavuddin A., and Hasan M., "DeepThin: A Novel Lightweight CNN Architecture for Traffic Sign Recognition without GPU Requirements," *Expert Systems with Applications*, vol. 168, pp. 114481, 2021. https://doi.org/10.1016/j.eswa.2020.114481

[14] Harizi R., Walha R., and Drira F., "Deep-Learning Based End-to-End System for Text Reading in the Wild," *Multimedia Tools and Applications*, vol. 81, no. 17, pp. 24691-24719, 2022. https://doi.org/10.1007/s11042-022-11998-x

[15] He X., Yuan J., Li M., Wang R., Wang H., and Li Z., "A Text-Specific Domain Adaptive Network for Scene Text Detection in the Wild," *Applied Intelligence*, vol. 53, no. 22, pp. 26827-26839, 2023. https://doi.org/10.1007/s10489-023-04873-1

[16] Jain M., Mathew M., and Jawahar C., "Unconstrained OCR for Urdu Using Deep CNN-RNN Hybrid Networks," *in Proceedings of the 4th IAPR Asian Conference on Pattern Recognition*, pp. 747-752, Nanjing, 2017. DOI:10.1109/ACPR.2017.5

[17] Jurisic F., Filkovic I., and Kalafatic Z., "Multiple-Dataset Traffic Sign Classification with OneCNN," *in Proceedings of 3rd Asian Conference on Pattern Recognition*, Kuala Lumpur, pp. 614-618, 2015. DOI:10.1109/ACPR.2015.7486576

[18] Karatzas D., Gomez-Bigorda L., Nicolaou A., Ghosh S., Bagdanov A., Iwamura M., Matas J., Neumann L., Chandrasekhar V., Lu S., Shafait F., Uchida S., and Valveny E., "ICDAR Competition on Robust Reading," *in Proceedings of the 13th International Conference on Document Analysis and Recognition*, Tunis, pp. 1156-1160, 2015. DOI:10.1109/ICDAR.2015.7333942.

[19] Karatzas D., Shafait F., Uchida S., Iwamura M., Bigorda L., Mestre S., Mas J., Mota D., Almazan J., and Heras L., "ICDAR Robust Reading Competition," *in Proceedings of the 12th International Conference on Document Analysis and Recognition*, Washington (DC), pp. 1484-1493, 2013. DOI:10.1109/ICDAR.2013.221

[20] Larsson F., Felsberg M., and Forssen P., "Correlating Fourier Descriptors of Local Patches for Road Sign Recognition," *IET Computer Vision*, vol. 5, no. 4, pp. 244-254, 2011. DOI:10.1049/iet-cvi.2010.0040

[21] Lazzeretti L., Innocenti N., Nannelli M., and Oliva S., "The Emergence of Artificial Intelligence in the Regional Sciences: A Literature Review," *European Planning Studies*, vol. 31, no. 7, pp. 1304-1324, 2023. https://doi.org/10.1080/09654313.2022.2101880

[22] Li J. and Wang Z., "Real-Time Traffic Sign Recognition Based on Efficient CNNs in the Wild," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 975-984, 2019. DOI:10.1109/TITS.2018.2843815

[23] Li J., Liang X., Wei Y., Xu T., Feng J., and Yan S., "Perceptual Generative Adversarial Networks for Small Object Detection," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, pp. 1951-1959, 2017. DOI:10.1109/CVPR.2017.211

[24] Li X., Song R., Fan J., Liu M., and Wang F., "Development and Testing of Advanced Driver Assistance Systems through Scenario-based System Engineering," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 8, pp. 3968-3973, 2023. DOI:10.1109/TIV.2023.3297168

[25] Li X., Wang W., Hou W., Liu R., Lu T., and Yang J., "Shape Robust Text Detection with Progressive Scale Expansion Network," *arXiv Preprint*, vol. arXiv:1806.02559, pp. 1-12, 2018. https://arxiv.org/pdf/1806.02559

[26] Liao M., Shi B., Bai X., Wang X., and Liu W., "TextBoxes: A Fast Text Detector with a Single Deep Neural Network," *in Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, pp. 4161-4167, 2017. https://dl.acm.org/doi/10.5555/3298023.3298172

[27] Lin H., Yang P., and Zhang F., "Review of Scene Text Detection and Recognition," *Archives of Computational Methods in Engineering*, vol. 27, no. 2, pp. 433-454, 2019. https://doi.org/10.1007/s11831-019-09315-1

[28] Lin T., Dollár P., Girshick R., He K., Hariharan B., and Belongie S., "Feature Pyramid Networks for Object Detection," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulum, pp. 2117-2125, 2017. DOI:10.1109/CVPR.2017.106

[29] Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C., and Berg A., "SSD: Single Shot MultiBox Detector," *in Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, pp. 21-37, 2016. https://link.springer.com/book/10.1007/978-3-319-46448-0

[30] Long J., Shelhamer E., and Darrell T., "Fully Convolutional Networks for Semantic

Segmentation," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, pp. 3431-3440, 2015. DOI:10.1109/CVPR.2015.7298965

[31] Ma D., Lin Q., and Zhang T., "Mobile Camera Based Text Detection and Translation," *Stanford University*, pp. 1-5, 2000. https://stacks.stanford.edu/file/druid:my512gb2187/Ma_Lin_Zhang_Mobile_text_recognition_and_translation.pdf

[32] Meng Z., Fan X., Chen X., Chen M., and Tong Y., "Detecting Small Signs from Large Images," *in Proceedings of the International Conference on Information Reuse and Integration*, San Diego, pp. 217-224, 2017. DOI:10.1109/IRI.2017.57

[33] Mogelmose A., Trivedi M., and Moeslund T., "Vision-Based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1484-1497, 2012. DOI:10.1109/TITS.2012.2209421

[34] Mohammad S., "Artificial Intelligence in Information Technology," *SSRN Electronic Journal*, pp. 1-15, 2020. http://dx.doi.org/10.2139/ssrn.3625444

[35] Nayef N., Yin F., Bizid I., Choi H., Feng Y., Karatzas D., Luo Z., Pal U., Rigaud C., and Chazalon J., "ICDAR Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification-RRC-MLT," *in Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition*, Kyoto, pp. 1454-1459, 2017. DOI:10.1109/ICDAR.2017.237

[36] Nirmala P., Ramesh S., Tamilselvi M., Ramkumar G., and Anitha G., "An Artificial Intelligence Enabled Smart Industrial Automation System Based on Internet of Things Assistance," *in Proceedings of the International Conference on Advances in Computing, Communication and Applied Informatics*, Chennai, pp. 1-6, 2022. DOI:10.1109/ACCAI53970.2022.9752651

[37] Panneerselvam J., Subramaniam B., and Meenakshisundaram M., "A Cognitive Approach to Predict the Multi-Directional Trajectory of Pedestrians," *The International Arab Journal of Information Technology*, vol. 20, no. 2, pp. 242-252, 2023. https://doi.org/10.34028/iajit/20/2/11

[38] Raisi Z. and Zelek J., "Text Detection and Recognition in the Wild for Robot Localization," *arXiv Preprint*, vol. arXiv:2205.08565v2, pp. 163-174, 2022. DOI:10.48550/arXiv.2205.08565

[39] Ramesh M. and Mahesh K., "A Performance Analysis of Pre-Trained Neural Network and Design of CNN for Sports Video Classification," *in Proceedings of the International Conference on Communication and Signal Processing*, Chennai, pp. 0213-0216, 2020. DOI:10.1109/ICCSP48568.2020.9182113

[40] Rawlley O. and Gupta S., "Artificial Intelligence-Empowered Vision-Based Self-Driver Assistance System for Internet of Autonomous Vehicles," *Transactions on Emerging Telecommunications Technologies*, vol. 34, no. 2, pp. e4683, 2023. https://doi.org/10.1002/ett.4683

[41] Redmon J. and Farhadi A., "Yolov3: An Incremental Improvement," *arXiv Preprint*, vol. arXiv:1804.02767, pp. 1-6, 2018. https://arxiv.org/pdf/1804.02767

[42] Redmon J., Divvala S., Girshick R., and Farhadi A., "You Only Look Once: Unified, Real-Time Object Detection," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 779-788, 2016. DOI:10.1109/CVPR.2016.91

[43] Ren S., He K., Girshick R., and Sun J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *in Proceedings of the 28th International Conference on Neural Information Processing Systems*, Montreal, pp. 91-99, 2015. https://dl.acm.org/doi/10.5555/2969239.2969250

[44] Rong X., Yi C., and Tian Y., "Recognizing Text-Based Traffic Guide Panels with Cascaded Localization Network," *in Proceedings of the ECCV Workshops*, Amsterdam, pp. 109-121, 2016. https://doi.org/10.1007/978-3-319-46604-0_8

[45] Schaefer S., McPhail T., and Warren J., "Image Deformation Using Moving Least Squares," *AMC Transitions on Graphics*, vol. 25, no. 3, pp. 533-540, 2006. https://doi.org/10.1145/1141911.1141920

[46] Sermanet P. and LeCun Y., "Traffic Sign Recognition with Multi-Scale Convolutional Networks," *in Proceedings of the International Joint Conference on Neural Networks*, San Jose, pp. 2809-2813, 2011. DOI:10.1109/IJCNN.2011.6033589

[47] Shahab A., Shafait F., and Dengel A., "ICDAR Robust Reading Competition Challenge 2: Reading Text in Scene Images," *in Proceedings of the International Conference on Document Analysis and Recognition*, Beijing, pp. 1491-1496, 2011. DOI:10.1109/ICDAR.2011.296

[48] Shaout A., Mysuru D., and Raghupathy K., "Vehicle Condition, Driver Behavior Analysis and Data Logging through CAN Sniffing," *The International Arab Journal of Information Technology*, vol. 16, no. 3A, pp. 493-498, 2019. https://ccis2k.org/iajit/PDF/Special%20Issue%202019,%20No.%203A/18594.pdf

[49] Shi B., Bai X., and Belongie S., "Detecting Oriented Text in Natural Images by Linking Segments," *in Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition, Honolulu, pp. 2550-2558, 2017. https://openaccess.thecvf.com/content_cvpr_2017/papers/Shi_Detecting_Oriented_Text_CVPR_2017_paper.pdf

[50] Shi X., Peng G., Shen X., and Zhang C., "TextFuse: Fusing Deep Scene Text Detection Models for Enhanced Performance," *Multimedia Tools and Applications*, vol. 83, pp. 22433-22454, 2024. https://doi.org/10.1007/s11042-023-16389-4

[51] Shorten C. and Khoshgoftaar T., "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1-48, 2019. https://doi.org/10.1186/s40537-019-0197-0

[52] Simonyan K. and Zisserman A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv Preprint*, vol. arXiv:1409.1556, pp. 1-14, 2015. https://doi.org/10.48550/arXiv.1409.1556

[53] Stallkamp J., Schlipsing M., Salmen J., and Igel C., "The German Traffic Sign Recognition Benchmark: A Multi-Class Classification Competition," *in Proceedings of the International Joint Conference on Neural Networks*, San Jose, pp. 1453-1460, 2011. DOI:10.1109/IJCNN.2011.6033395

[54] Sun Q., Xiao Z., and Ji P., "Improved CTPN Based Attention Mechanism for Scene Text Detection," *in Proceedings of the 2nd International Conference on Big Data, Artificial Intelligence and Risk Management*, Xian, pp. 199-202, 2022. DOI:10.1109/ICBAR58199.2022.00045

[55] Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., and Rabinovich A., "Going Deeper with Convolutions," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, pp. 1-9, 2015. DOI:10.1109/CVPR.2015.7298594

[56] Tian Z., Huang W., He T., He P., and Qiao Y., "Detecting Text in Natural Image with Connectionist Text Proposal Network," *in Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, pp. 56-72, 2016. https://link.springer.com/chapter/10.1007/978-3-319-46484-8_4

[57] Timofte R., Zimmermann K., and Van Gool L., "Multi-View Traffic Sign Detection, Recognition, and 3D Localisation," *in Proceedings of the Workshop on Applications of Computer Vision*, Snowbird, pp. 633-647, 2014. DOI:10.1109/WACV.2009.5403121

[58] Tounsi M., Moalla I., and Alimi A., "ARASTI: A Database for Arabic Scene Text Recognition," *in Proceedings of the 1st International Workshop on Arabic Script Analysis and Recognition*, Nancy, pp. 140-144, 2017. DOI:10.1109/ASAR.2017.8067776

[59] Turki H., Elleuch M., Kherallah M., Syphax Traffic Panels Dataset, IEEE Dataport, https://dx.doi.org/10.21227/5zd9-pe55, Last Visited, 2024.

[60] Turki H., Halima M., and Alimi A., "Text Detection Based on MSER and CNN Features," *in Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition*, Kyoto, pp. 949-954, 2017. DOI:10.1109/ICDAR.2017.159

[61] Veit A., Matera T., Neumann L., Matas J., and Belongie S., "COCO Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images," *arXiv Preprint*, arXiv:1601.07140, pp. 1-8, 2016. https://arxiv.org/pdf/1601.07140

[62] Wada K., Labelme: Image Polygonal Annotation with Python, Github, 2016, https://github.com/labelmeai/labelme, Last Visited, 2024.

[63] Wan S., Ding S., and Chen C., "Edge Computing Enabled Video Segmentation for Real-Time Traffic Monitoring in Internet of Vehicles," *Pattern Recognition*, vol. 121, pp. 108146, 2022. https://doi.org/10.1016/j.patcog.2021.108146

[64] Wang J., Chen Y., Dong Z., and Gao M., "Improved YOLOv5 Network for Real-Time Multi-Scale Traffic Sign Detection," *Neural Computing and Applications*, vol. 35, no. 10, pp. 7853-7865, 2023. https://link.springer.com/article/10.1007/s00521-022-08077-5

[65] Wang K. and Belongie S., "Word Spotting in the Wild," *in Proceedings of the 11th European Conference on Computer Vision*, Heraklion, pp. 591-604, 2010. https://link.springer.com/chapter/10.1007/978-3-642-15549-9_43

[66] Wikipedia, https://en.wikipedia.org/wiki/Sfax, Last Visited, 2024.

[67] Wu Y., Li Z., Chen Y., Nai K., and Yuan J., "Real-Time Traffic Sign Detection and Classification towards Real Traffic Scene" *Multimedia Tools and Applications*, vol. 79, no. 25, pp. 18201-18219, 2020. https://doi.org/10.1007/s11042-020-08722-y

[68] Yousfi S., Berrani S., and Garcia C., "ALIF: A Dataset for Arabic Embedded Text Recognition in TV Broadcast," *in Proceedings of the 13th International Conference on Document Analysis and Recognition*, Tunis, pp. 1221-1225, 2015. DOI:10.1109/ICDAR.2015.7333958

[69] Zayene O., Hennebert J., Touj S., Ingold R., and Amara N., "A Dataset for Arabic Text Detection, Tracking and Recognition in News Videos-AcTiV," *in Proceedings of the 13th International Conference on Document Analysis and* Recognition, Tunis, pp. 996-1000, 2015. DOI:10.1109/ICDAR.2015.7333911

[70] Zeng W., Meng Q., and Zhang S., "Natural Scene

Chinese Character Text Detection Method Based on Improved CTPN," *in Proceedings of the 3rd International Conference on Electrical, Mechanical and Computer Engineering*, Guizhou, vol. 1314, no. 1, pp. 1-7, 2019. DOI:10.1088/1742-6596/1314/1/012200

[71] Zhang H., Zhao K., Song Y., and Guo J., "Text Extraction from Natural Scene Image: A Survey," *Neurocomputing*, vol. 122, pp. 310-323, 2013. https://doi.org/10.1016/j.neucom.2013.05.037

[72] Zhang J., Xie Z., Sun J., Zou X., and Wang J., "A Cascaded R-CNN with Multiscale Attention and Imbalanced Samples for Traffic Sign Detection," *IEEE Access*, vol. 8, pp. 29742-29754, 2020. DOI:10.1109/ACCESS.2020.2972338

[73] Zhang Q., Zhang M., Chen T., Sun Z., Ma Y., and Yu B., "Recent Advances in Convolutional Neural Network Acceleration," *Neurocomputing*, vol. 323, pp. 37-51, 2019. https://doi.org/10.1016/j.neucom.2018.09.038

[74] Zhong L., Zheng X., and Su Y., "Improved EAST Scene Text Detection Based on ResNet-50," *in Proceedings of the International Conference on Computer Vision, Application, and Algorithm*, Chongqing, pp. 155-159, 2022. https://doi.org/10.1117/12.2673275

[75] Zhou X., Yao C., Wen H., Wang Y., Zhou S., He W., and Liang J., "East: An Efficient and Accurate Scene Text Detector," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, pp. 5551-5560, 2017. DOI:10.1109/CVPR.2017.283

[76] Zhu Y., Zhang C., Zhou D., Wang X., Bai X., and Liu W., "Traffic Sign Detection and Recognition Using Fully Convolutional Network Guided Proposals," *Neurocomputing*, vol. 214, pp. 758-766, 2016. https://doi.org/10.1016/j.neucom.2016.07.009

[77] Zhu Z., Liang D., Zhang S., Huang X., Li B., and Hu S., "Traffic-Sign Detection and Classification in the Wild," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 2110-2118, 2016. DOI:10.1109/CVPR.2016.232

Houssem Turki received the B.S. degree in computer science from The Faculty of Economics and Management of Sfax-Tunisia (FSEGS) and completed his master's degree in New Technologies of Dedicated Computer Systems from the National School of Engineering of Sfax (ENIS), Tunisia. His main research concerns object/text Detection in Natural Scene Image.

Mohamed Elleuch was born in Sfax, Tunisia. He received the Ph.D. degree at the National School of Computer Science (ENSI), Manouba-Tunisia in 2017, from University of Manouba. Now he is Assistant Professor in the Higher Institute of Computing and Management of Kairouan (ISIGK), University of Kairouan. The research topic of his studies was pattern recognition based on Artificial Intelligence Machine Learning and Deep Learning algorithms. The applications are focused on Handwriting Text and Plant Diseases Recognition, Medical Image Processing, etc. He is reviewer of several international journals.

Kamal Othman received the Ph.D. degree from Simon Fraser University, Burnaby, BC, Canada, in 2020. He joined with the Department of Electrical Engineering, Umm Al-Qura University Makkah, Saudi Arabia, where he is currently an Assistant Professor. His research interests include Applied Artificial Intelligence, Deep Learning, and Robotics.

Monji Kherallah graduated in Electrical Engineering 1989, obtained a Ph.D. in Electrical Engineering in 2008. He is now a professor in Electrical and Computer Engineering at the University of Sfax. His research interest includes applications of Intelligent Methods to Pattern Recognition and Industrial Processes. He focuses his research on Handwritten Documents Analysis and Recognition, online Arabic Handwriting Recognition, pattern recognition and image processing. He is a reviewer of several international journals.