Multimedia Tools and Applications (2023) 82:34793–34808
https://doi.org/10.1007/s11042-023-15062-0

PESTD: a large-scale Persian-English scene text dataset

Atefeh Ranjkesh Rashtehroudi[1] · Alireza Akoushideh[2] · Asadollah Shahbahrami[1]

Received: 23 April 2022 / Revised: 5 November 2022 / Accepted: 2 March 2023 /
Published online: 25 March 2023

Abstract

Extracting text from natural scene images has become a vital issue. The uncertainty of size, color, background, and alignment of the characters make text recognition in natural scene images a demanding challenge. Also, another recent challenge has been the development and expansion of intelligent systems in the field of transportation, especially the recognition of traffic signs, which help ensure safer and easier driving. Therefore, existing a scene-text dataset as a benchmark to generalize researchers' algorithms is critical. This study, as one of the first studies in the field of text-based traffic signs, intends to prepare a Persian-English multilingual dataset (PESTD) that includes 5832 instances including letters, digits, and symbols in three categories: Persian, English, and Persian-English. Due to the similarity of the calligraphy of numbers and letters in Persian (Farsi), Arabic and Urdu languages, The PESTD can be used in all countries with these languages. To prepare PESTD instances, the text detection process was performed on the traffic signs in Iran. The CRAFT feature extraction algorithm with YOLO and the Tesseract engine have been combined to take an effective step to recognize cursive and multilingual languages despite their specific challenges. Experimental results depict that the values of the evaluation criteria in YOLOv5 are better than its older versions. The accuracy and F1-score values on the PESTD have been attained at 95.3% and 92.3%, respectively.

Keywords Cursive script · Deep learning · Farsi/Arabic/Urdu · Farsi-English · Multilingual · Persian-English · Scene text dataset

✉ Alireza Akoushideh
akushide@tvu.ac.ir

[1] Computer Engineering Department, Guilan University, Rasht, Iran

[2] Electrical and Computer Department, Technical and Vocational University (TVU), Guilan Branch, Rasht, Iran

1 Introduction

Scene images are used based on their application in different situations and to achieve different goals. Therefore, the detection of them in various fields such as Industrial automation [6], Robot navigation [26], Intelligent transportation system (ITS) [40], Search and translation [8, 31], Optical character recognition (OCR) [41], and other computer vision applications will lead to the extraction of very useful and efficient information from them. Intelligent Transportation Systems (ITSs) need effective features for performing their algorithms. Automatic detection of traffic signs is a critical component of an advanced driver assistance system (ADAS), and in future vehicles, it will be an integral component [9]. Therefore, designing an intelligent system can dramatically assist drivers and significantly reduce the rate of traffic accidents. Nonetheless, with the increasing demand for smart vehicles, the automatic and online detection and recognition of traffic signs is vital, a task that can be facilitated with computer vision.

Scene text detection is a type of text area detection in a complex background. The scene text is usually seen in various fonts and sizes and often with a background in urban environments with various noises, making scientific investigation difficult [20]. The application of text-based traffic signs in ITSs is one of the most important and widely used subsets of scene images. The extensive potential applications of this field of computer vision can lead to multiple challenges that can be generally classified into the complexity of background, diversity of scene text, and interference factors (noise) [41]. Noise in Scene text is one of the challenges that affect text detection. Since the images are taken from natural environments, factors such as light intensity, text angle, color, dust, and tree branches can affect image quality.

In this regard, the scene-text datasets with a wide variety of instances help the researchers to generalize their vision-based algorithms [30]. Given the expansion of Internet communication and the existence of multilingual countries or organizational and academic documents, here, the authors focused on preparing a collection of bilingual Persian/Arabic and English datasets with sufficient sample quantity and diversity to recognize text-based images [17]. Our contributions to this paper are the following:

- Proposing a complete and large-scale dataset of Persian-English named PESTD at the word level. PESTD includes 5832 instances including letters, digits, and symbols in three categories: Persian, English, and Persian-English.

Note that all Arabic and Persian numbers are similar. Also, the Arabic letters are similar to 28 out of 32 Persian letters. In addition, the word “Farsi” is an alternative name for “Persian”. Therefore, the PESTD can be used in all countries whose official language is Farsi, Arabic, or Urdu. On the other hand, because this data set is bilingual, its English samples can be used in many countries.

- To prepare PESTD instances, the text detection process was performed on the traffic signs in Iran. The CRAFT feature extraction algorithm with YOLO and the Tesseract engine have been combined to take an effective step to recognize cursive and multilingual languages despite their specific challenges. The use of an end-to-end structure has made this architecture usable in other applications and research. For example, the detection model of the proposed idea can be used to detect texts in other applications such as manuscripts and typed texts with different fonts.

- The proposed data set includes six general categories of challenges such as weather conditions, light intensities conditions, distance, background (surrounding environment such as trees, streets, cars, buildings, …, and board structure), color, and view angle. This issue causes the performance of the proposed method to be evaluated comprehensively.
- The accuracy and F-score values (evaluation criteria in YOLOv5) on the PESTD have been attained at 95.3% and 92.3%, respectively.

The rest of this paper is organized as follows: Section 2 covers several types of research related to scene-text detection and related datasets. Section 3 discusses the method used to prepare the proposed dataset and characterizes it. Section 4 examines the dataset efficiency, and Section 5 evaluates the introduced dataset against other datasets. Finally, the conclusions and future research suggestions are presented in Section 6.

2 Related work

As mentioned before, the various applications of text detection and recognition have turned them into curious topics in computer vision. In general, the scene-text datasets are divided into three categories: handwritten, printed, and scene text based images. Each category is classified into two groups of real and synthetic datasets depending on the data collection method. Real datasets are created by scanning documents (e.g., newspapers and journals) or scene images. However, existing texts are used to construct synthetic datasets. An image of each character or word is created randomly by applying various fonts and sometimes various backgrounds. More images can be produced using the existing images.

This article focuses on preparing bilingual Persian-English datasets of scene text images. This section thus investigates only the scene text datasets. Some of the most well-known English real datasets are SynthText [10], Synth90k [12], and VerisimilarSynthesis [39]. Among these datasets, due to the approximate similarity of Arabic and Persian (or Farsi) languages, two synthetic datasets, ACTIV [38] and ALIF [37], can be referred which have been extracted from video frames of Arabic channels. The ALIF dataset is larger than the ACTIV dataset. Each dataset contains 6532 images of text and 21,520 images of words from Arabic channels, respectively. In the following, different types of datasets are introduced:

- *Real dataset:* The real datasets in the scene text images are divided into three categories: regular, irregular, and multilingual.
- *Regular dataset:* The most famous regular datasets in English are ICDAR 2003 (IC03) [16], ICDAR 2013 (IC13) [13], IIIT 5 k-word (IIIT5k) [18], and Street View Text (SVT) [34]. This includes a test dataset containing 251 scene images with labeled text bounding boxes, 1015 ground truths cropped word images, 3000 cropped word test images collected from the Internet, and 249 street view images collected from Google Street View, respectively. In this (real dataset) category for the Arabic language, there are two datasets: ARASTI [29] and ARASTEC [28]. They include 1687 images, 1280 isolated Arabic words, 2093 isolated Arabic letters, and 60 scene text images. In this case, a slight comparison indicates a highly nonsignificant variety in this category of datasets for Arabic and Persian (or Farsi).
- *Irregular dataset:* In irregular datasets, most text samples have a low resolution with different fonts that are not written horizontally but are in curved format, causing this

dataset to face more challenges than other categories. As an irregular English dataset, the COCO-Text [32] contains no-text, legible, and illegible text images. In total, there are 22,184 training images and 7026 validation images with at least one sample of legible text. The ICDAR 2015 dataset (IC15) [14] contains 1500 images, 1000 for training and 500 for testing. The StreetViewText-Perspective (SVT-P) dataset [21] contains 238 images with 639 cropped text instances.

- *Multilingual datasets:* The frequency of bilingual and multilingual texts directly relates to urban development. Ahmad et al. [1] introduce a bilingual English-Arabic dataset. Table 1 compares these datasets. Note that the number of multilingual scene-text datasets is limited. In addition, there are no rich Persian-English multilingual datasets, whether in the form of prints or scene text images. Therefore, this research attempts to tackle this problem for the first time by preparing a collection of bilingual Persian-English datasets.

3 Proposed dataset

3.1 Description

Traffic signs are generally divided into three several categories. The first category includes *circular regulatory signs* with a red border that indicate the type of prohibition by a special sign. Among the stop signs, the stop sign has been considered as an octagon with a completely red background and with the word STOP written for the importance and accuracy of the drivers. Due to the importance of the sign of observing the right of precedence, only the top of this sign is downwards. The second category includes *warning* signs, mainly in the form of a triangle with red border stripes and white background, one end of which is upwards, and inside which the type of danger is indicated by special black markings. The third category includes guide signs that contain general advice. These are designed in white, green, brown, yellow, and blue, and triangle, circle, rectangle, and square shapes. Figure 1 depicts three different types of traffic signs.

In addition to the mentioned traffic sign categories, there is another type of signage called *path guide* signs (urban route guidance (Fig. 2). They contain much more text than the three categories introduced. These signs are generally rectangular and can only be designed as flags at the exits (with a sharp arrow-like ending to indicate a specific direction). Traffic signs convey some important guiding and destination information to drivers and pedestrians. This information includes transit conditions, facilities, and access to the route. In some cases, these signs also include regulatory orders. The particular line of these signs should be such that it conveys messages to all drivers quickly and efficiently. Therefore, in the design of the font, in addition to the appropriate size, readability must also be taken into account. Only one font should be selected as the standard font and used in all signs. Currently, two fonts have been selected: Gem for Persian texts and Homa for English texts in urban route signs. In some cases, the Abrisham font is also used. The size of the text in the signs is a function of the time required to read the text. This time depends on the speed of the vehicle approaching the sign. Determining the text size is especially important for route signs. The size of the text, which is measured by the height of the mosaic of Persian letters, is a function of several parameters, such as the number of words, speed of movement, and the distance of the board from the axis

Table 1 Description of the scene-text benchmark datasets

Datasets	Language	Data Distribution Details						Label		Type
		Images	Training Pictures	Testing Pictures	Instances	Training Instances	Testing Instances	Char	Word	
SynthText [10]	English	~6,000,000	N/A	N/A	~6,000,000	N/A	N/A	√	√	Synthetic, Regular
Synth90k [12]		~9,000,000	N/A	N/A	~9,000,000	N/A	N/A	×	√	
Verisimilar Synthesis [39]		N/A	N/A	N/A	~5000,000	N/A	N/A	×	√	
ACTIV [38]	Arabic	21,520	N/A	N/A	N/A	N/A	N/A	×	√	
ALIF [37]		6532	N/A	N/A	5052	4152	900	√	√	
IC03 [16]	English	509	258	251	2268	1157	1111	√	√	Real, Regular
IC13 [13]		561	420	141	5003	3564	1439	√	√	
IIIT5k [18]		1120	380	740	5000	2000	3000	√	√	
SVT [34]		350	100	250	725	211	514	×	√	
ARASTI [29]	Arabic	1280	N/A	N/A	374	N/A	N/A	√	√	
ARASTEC [28]		60	N/A	N/A	N/A	N/A	N/A	√	√	
COCO-Text [32]	English	63,686	43,686	10,000	145,859	118,309	27,550	×	√	Real, Irregular
IC15 [14]		1500	1000	500	6545	4468	2077	×	√	
SVT-P [21]		238	0	238	639	0	639	×	√	
Ahmad et.al [1]	English-Arabic	2469	N/A	N/A	N/A	N/A	N/A	×	√	Real, Multilingual

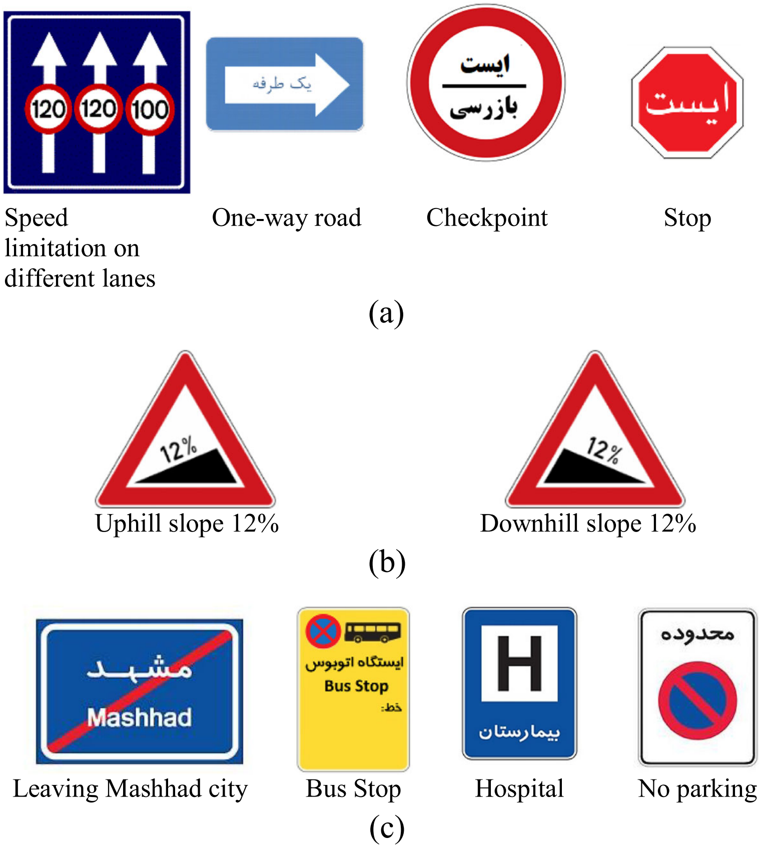
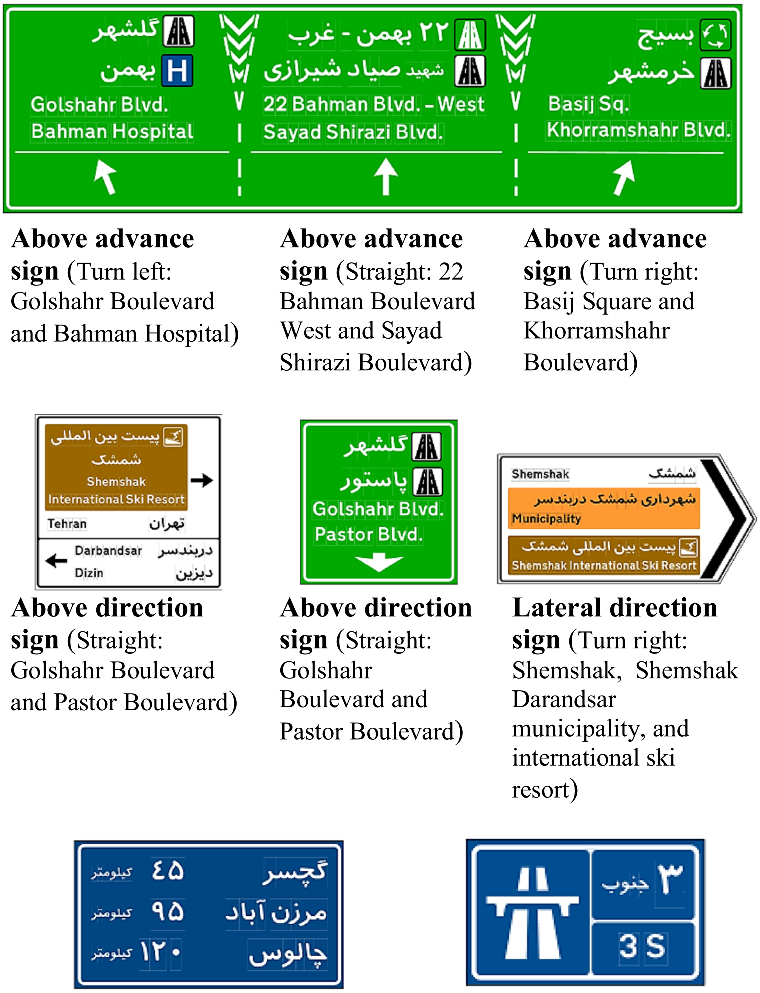


Fig. 1 The examples of three different types of traffic signs (**a**) Circular regulatory signs, (**b**) Warning signs, (**c**) Path Guide signs

of view. The size of the direction signs is determined based on the height of the mosaic of Persian letters, the volume of the text, and the arrangement of other elements. The height of the letters is also determined based on design speed. Signs are used to make an intended message more expressive and accelerate the comprehension of the text. Route signs should be addressed at specific intervals depending on the conditions, and type of road, before the location and finally installed at the entrance to the access road. Except in exceptional cases, signs should be installed on the right side of the path.

3.2 Proposed method for data collection

There are many monolingual countries where there are also traces of other languages. The official language of Iran is Persian (Farsi), but in many cases, such as universities, organizations, scene images, websites, or similar cases, a combination of English and Arabic with Persian is used. Accordingly, a complete and exhaustive dataset is needed in the first stage to recognize the multilingual texts in the images. Since text recognition is preceded by text detection, accurate detection of the text area is an essential step for improving recognition accuracy and reducing the computational load due to processing the text area instead of the



Above advance sign (Turn left: Golshahr Boulevard and Bahman Hospital)

Above advance sign (Straight: 22 Bahman Boulevard West and Sayad Shirazi Boulevard)

Above advance sign (Turn right: Basij Square and Khorramshahr Boulevard)

Above direction sign (Straight: Golshahr Boulevard and Pastor Boulevard)

Above direction sign (Straight: Golshahr Boulevard and Pastor Boulevard)

Lateral direction sign (Turn right: Shemshak, Shemshak Darandsar municipality, and international ski resort)

Distance to Gachsar 45 km, Marzan Abad 95km and Chalus 120 km

Freeway No. 3 to the south

Fig. 2 The examples of Path confirmation signs

whole image. Accordingly, here the authors aim to prepare a dataset by taking into account the detection and recognition phases separately. In this research, the Character Region Awareness for Text detection (CRAFT) model has been used to produce a dataset for Persian-English scene text. The backbone of the CRAFT model's feature extraction architecture is based on the VGG-16 network architecture. This includes region and affinity for giving the character region [2] and the affinity of the characters to combine. Therefore, with this approach, the texts in the image are first processed at the character level, and in the next step, they combine according to the affinity score and form the word (Fig. 3).

By applying CRAFT for each sample, the text area is detected and extracted from the base image (Fig. 4). Here, the CRAFT has been applied to different samples with different challenges. The results show that CRAFT performs desirably in detecting text areas in different

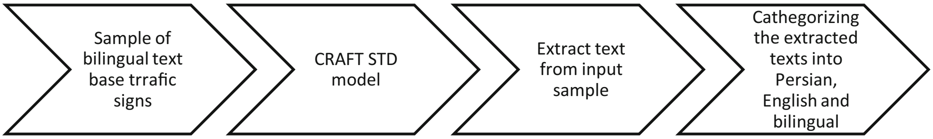

Sample of bilingual text base trrafic signs
CRAFT STD model
Extract text from input sample
Cathegorizing the extracted texts into Persian, English and bilingual

Fig. 3 The scene text database creating flow

conditions, such as distance from the traffic signs (far or near) (Fig. 4-b), its height from the ground, background color, shape of the sign, amount of light in space, and location. In addition, in more complex situations, when there are multiple texts in the sample, such as license plates and traffic signs (Fig. 4-a and Fig. 4-c), this system can optimally carry out the detection operation. Although the samples in the signs have neat text in the range of font and size, the CRAFT model was also able to recognize the handwritten text in the image, in addition to the text written on the sign, with desirable accuracy.

The criterion for evaluating the performance of the CRAFT model on the dataset used is calculated by two precision and recall metrics based on the following equations:

$$precision = \frac{\sum_{i=1}^{|D|} matchD(D_i)}{|D|} \tag{1}$$

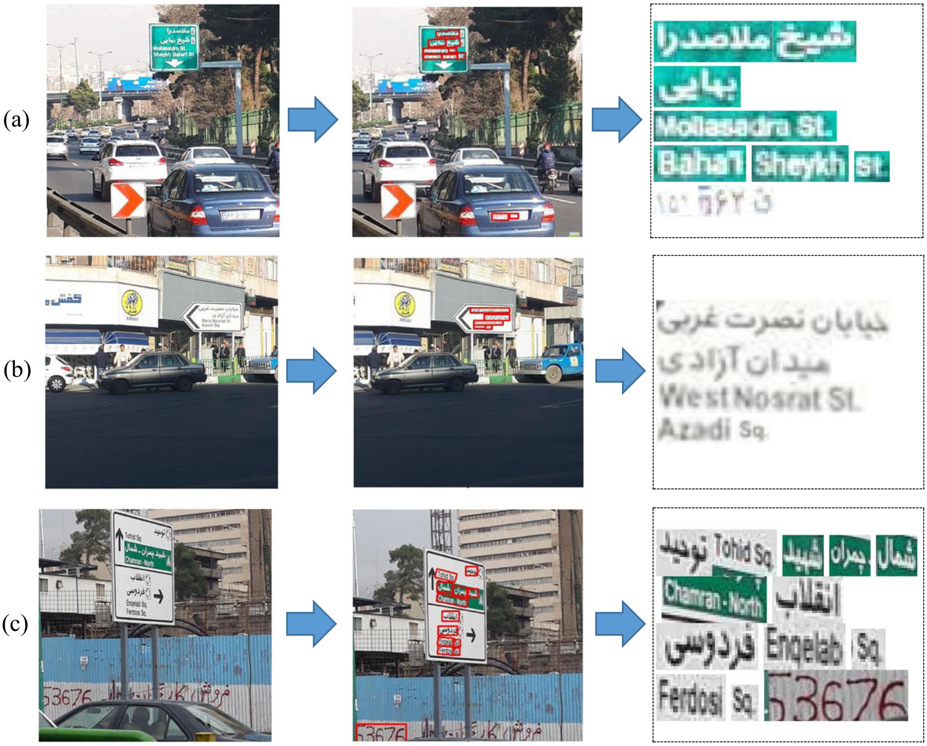


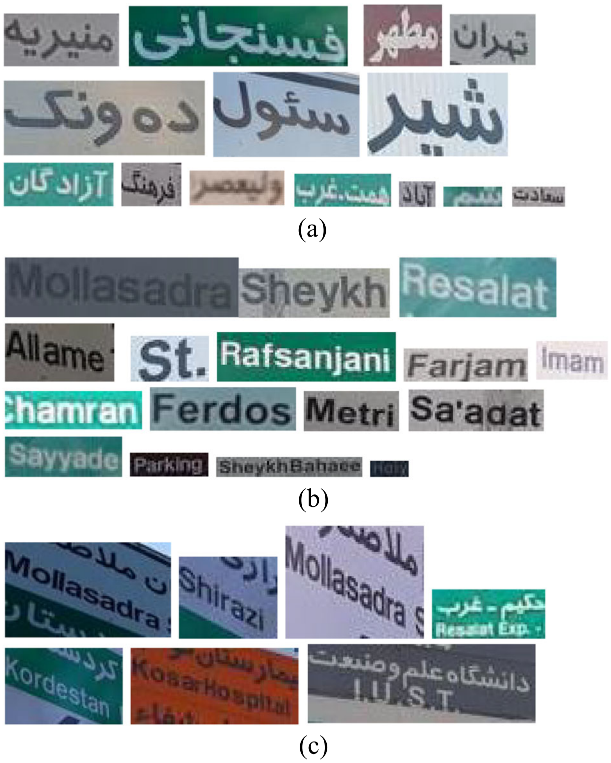
Fig. 4 The results of the CRAFT text detection model. (**a**) License plate detection, (**b**) Text detection; Scale challenge, and (**c**) Handwriting text detection and slant challenge

$$recall = \frac{\sum_{i=1}^{|G|} matchG(G_i)}{|G|} \tag{2}$$

Where D is the list of detected rectangles and G is the list of ground-truth rectangles. Various types of matching functions (matchG and matchD) exist between ground truth and detected rectangles, such as one-to-one, one-to-many, and many-to-one [35]. Based on the one-to-many matching function, the values obtained for precision and recall were 0.9705 and 0.9822. The prepared dataset is the first dataset of Persian-English scene test images prepared with the help of text-based traffic signs in Tehran. It can help solve a significant research problem that emanates from the lack of sufficient exhaustive text datasets in Persian, Arabic, Urdu, or similar languages. Because this dataset is the basis for the recognition of Persian texts in another study by us, so to promote research in Persian/Arabic, the prepared dataset will be publicly available for all studies of other researchers [Link]. Figure 5 depicts some instances in three categories: Persian, Persian-English, and English.

3.3 Dataset specifications

A text-based traffic sign dataset named “The Persian Text-Based Traffic Signs Dataset” was used in this study. It has 2643 instances containing Persian-English text as a basic dataset [15]. However, the prepared dataset recognizes and extracts Persian and English texts from the basic



(a)

(b)

(c)

Fig. 5 PESTD instances in three categories. (**a**) Persian, (**b**) English, and (**c**) Persian-English or Bilingual

dataset. The Persian texts in our dataset contain all 32 letters of the Persian alphabet, but in Persian or Arabic texts, the words are purely cursive, giving the letters of a word different shapes depending on their position (beginning, middle, end) in the word. According to the case mentioned above and other cases shown in Table 2, there are 122 different writing shapes according to Deutsches Institut für Normung (DIN) and International Phonetic Association (IPA) [11] standards for 32 Persian letters. The prepared dataset includes 5832 Persian and English words and numbers. Specifications of the proposed dataset (PESTD) have been mentioned in Table 3.

4 Experimental results

To take a step toward recognizing the introduced dataset, the single-stage deep learning technique of YOLO [25] version 3 and its combination with the Tesseract engine introduced in [23] have been used. In addition, an improved version of YOLO has been used as YOLOv4 [3] and YOLOv5 [36], so this improvement happened by having an improvement in the mean Average Precision (mAP) [33]. In YOLOv3 for the object detection step, Darknet-53 is used as a convolutional neural network (CNN), while in YOLOv4 and YOLOv5, the CSPdarkent53 has been used to act as a backbone [27].

Table 2 The Persian alphabet and their written shapes in the study dataset

row	Name	DIN 31635	IPA	Shapes			
				beginning	middle	end	single
1	همزه	ʾ	[ʔ]	ئـ	ـئـ	ـئ ـأ ـؤ	ء أ
2	الف	*ā*	[ɒ]	ـا			ا/آ
3	ب	*b*	[b]	بـ	ـبـ	ـب	ب
4	پ	*p*	[p]	پـ	ـپـ	ـپ	پ
5	ت	*t*	[t]	تـ	ـتـ	ـت	ت
6	ث	*s̱*	[s]	ثـ	ـثـ	ـث	ث
7	جیم	*j*	[d͡ʒ]	جـ	ـجـ	ـج	ج
8	چ	*č*	[t͡ʃ]	چـ	ـچـ	ـچ	چ
9	ح	*ḥ*	[h]	حـ	ـحـ	ـح	ح
10	خ	*x*	[x]	خـ	ـخـ	ـخ	خ
11	دال	*d*	[d]	ـد			د
12	ذال	*ẕ*	[z]	ـذ			ذ
13	ر	*r*	[ɾ]	ـر			ر
14	ز	*z*	[z]	ـز			ز
15	ژ	*ž*	[ʒ]	ـژ			ژ
16	سین	*s*	[s]	سـ	ـسـ	ـس	س
17	شین	*š*	[ʃ]	شـ	ـشـ	ـش	ش
18	صاد	*ṣ*	[s]	صـ	ـصـ	ـص	ص
19	ضاد	*z̧*	[z]	ضـ	ـضـ	ـض	ض
20	طا	*ṭ*	[t]	طـ	ـطـ	ـط	ط
21	ظا	*ẓ*	[z]	ظـ	ـظـ	ـظ	ظ
22	عین	ʿ	[ʕ]	عـ	ـعـ	ـع	ع
23	غین	*ġ*	[ɢ] / [ɣ]	غـ	ـغـ	ـغ	غ
24	ف	*f*	[f]	فـ	ـفـ	ـف	ف
25	قاف	*q*	[ɢ] / [ɣ] / [q] (in some dialects)	قـ	ـقـ	ـق	ق
26	کاف	*k*	[k]	کـ	ـکـ	ـک	ک
27	گاف	*g*	[ɡ]	گـ	ـگـ	ـگ	گ
28	لام	*l*	[l]	لـ	ـلـ	ـل	ل
29	میم	*m*	[m]	مـ	ـمـ	ـم	م
30	نون	*n*	[n]	نـ	ـنـ	ـن	ن
31	واو	*v* / *ū* / *ow*	[v] / [uː] / [o] / [ow] / [oː] (in Dari)	ـو			و
32	ه	*h*	[h]	هـ	ـهـ	ـه	ه
33	ی	*y* / *ī* / *á*	[j] / [i] / [ɒː] / [eː] (in Dari)	یـ	ـیـ	ـی	ی

Table 3 Samples of Persian/English dataset (PESTD) category details

Language	Samples numbers	Shapes number	Type	Instances
Persian	2614	122	Letter	See Table 1
		10	digits	**0** 9 8 7 6 5 4 3 2 1
		2	Symbol	◌ّ
English	3162	52	Letter	A-Z, a-z
		10	digits	0 1 2 3 4 5 6 7 8 9
		3	Symbol	- ‘.
Persian-English	56	199	all	All
Sum	**5862**	**255**		

The comparison of the YOLOv3, YOLOv4, and YOLOv5 methods on the PESTD (including 199 different forms of letters, numbers, and symbols in Persian and English) is shown in Table 4. Accuracy and F1-score as the criteria have been used to compare the YOLOv3, YOLOv4, and YOLOv5 in the detection step of the text recognition using the Tesseract engine. Accuracy is calculated from the ratio of the number of correct predictions to the total number of predictions. This is while the F1-score is a kind of averaging of precision and recall (as in Eq. (3) and Eq. (6), True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)) [22]. Both of these criteria are defined in the range of zero and one. So, their highest possible value is 1, and the lowest possible value is 0. For example, in the criterion of F1-score, value 1 indicates perfect precision and recall, and 0 indicates either the precision or recall is zero. In addition, the PC speed per second to evaluate the inference speed of algorithms to process data has been considered. The results show that YOLOv3 is less accurate than the other two versions because their backbone is different, and on the other hand, the performance of YOLOv5 is better than YOLOv4 due to the use of auto-learning bounding boxes [19] (Table 4). This is while the accuracy of the method mentioned in our previous study using YOLOv3 for the isolated Iranian license plate (including 27 different forms of letters and numbers in Persian) was almost 99%.

$$F1\text{-}score = 2 \times \frac{p \times r}{p + r} \tag{3}$$

Where, p and r are defined by Eq. (4), and Eq. (5).

$$p = \frac{TP}{TP + FP} \tag{4}$$

Table 4 Comparison of different detection methods for text recognition on PESTD

Methods	Accuracy (%)	F1-score (%)	Time (s)
YOLOv3 [24]	92	87.3	61.2
YOLOv4 [33]	94.1	91.1	56.1
YOLOv5 [36]	95.3	92.3	54

$$r = \frac{TP}{TP + FN} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

A personal computer with the following specifications has been considered for implementing the proposed algorithm:

- CPU: 8th Gen. Intel® Core™ i7–1.80GHz Processor.
- GPU: NVIDIA® GeForce® RTX 2070 SUPER ™ Turing ™ architecture with 8 GB GDDR6.
- RAM: 32 GB DDR4.
- Storage: 1 TB NVMe SSD.
- Operating System: UBUNTU 18.04.6.

5 Discussion

Since the dataset prepared in this paper was extracted from the scene text images of traffic signs based on Persian-English, its inherent features may challenge its users, especially in Persian. In the following, some challenges of the Persian scene text dataset and its limitations have been mentioned.

5.1 Challenges of Persian scene text dataset and its limitations

- Persian texts, unlike English texts, are written from right to left.
- The letters of words in Persian are often written cursively and, in some cases, non-separately, while in English, the letters are always written separately. Therefore, additional studies on the Persian language are relatively more complicated than in English. Therefore, according to the different shapes of the Persian alphabet in Table 2, this condition exists in 100% of the Persian alphabet.
- Some letters have the same shape. Letters can be distinguished from one another only based on the absence or presence of a dot (“ ” vs. “ ”), the number of dots (“ﺗ” vs. “ﺛ”), or the position of the dots (“ﭘ” vs. “ﺛ”) beneath or above letters. Furthermore, two Persian letters, “ ” and “ ” differ only in the existence of a stroke. Based on this, the only alphabets “واو“ ,” ـمـ“ ,”لامـ” and “ه” are not similar to other alphabets in any of the shapes that are placed at the beginning, middle, end, or single. Therefore, about 87.87% of all alphabets are slightly different from other alphabets in at least one shape.
- Most Persian letters (especially cursive letters) are jagged, and in cases where the image has noise, the jagged format of the character may be seen with the same baseline, complicating the recognition operation. This challenge is very dependent on the writing font, so a detailed analysis cannot be done on it.
- The letters in a word may overlap, meaning that a vertical line cannot wholly separate the letters. This challenge, like the previous challenge, depends on the font.

Table 5 Comparison of multilingual datasets

Dataset	Language	Font Style	Font Size	Image
The Maurdor project [4]	Handwritten and printed in French, English, and Arabic	Not disclosed	Not disclosed	5000 French text images, 2500 English text images, and 2500 Arabic text images
ALTID [7]	Handwritten and printed Arabic/English	Not disclosed	Not disclosed	Printed: 1845 Arabic text-block images and 2328 English text-block Handwritten: 460 Arabic and 582 English text-block images
MIDV-LAIT [5]	Printed Farsi, Arabic, Thai, and Indian	Normal, Bold, Italic	Multi sizes	3600 images
PESTD (ours)	Printed Farsi, Arabic, and English	Homa, Abrisham	Multi sizes	5832 images

5.2 Limitations

Since the motivation of this research is to provide a suitable dataset to create a context for detecting Persian (Farsi)/Arabic and English multilingual texts, these challenges can represent the difficulty of research in this field and demonstrate the value of the work in this field. However, to further evaluate this dataset, its font style and size and the number of samples were compared with other multilingual datasets (Table 5). The results show that the proposed dataset is more extensive in the number of samples, allowing for larger-scale detection of data with a richer variety. It should be noted that the proposed dataset includes samples under different illuminations at different angles and sizes but excludes all types of fonts and sizes.

6 Conclusion and future work

In this study, we present a bilingual Persian-English dataset (PESTD) based on the images of the traffic sign scenes that include 5832 instances including letters, digits, and symbols. The Persian texts in the dataset contain all 32 letters of the Persian alphabet with 122 different writing shapes according to the DIN and IPA. The instances in the presented dataset have been classified into Persian, Persian-English, and English categories. Regarding the similarity of Arabic, Persian, and Urdu numbers and letters, this dataset can be considered a suitable database in all regions with these languages. As a different challenge in comparison with the English language, the letters in this dataset are often written cursively and, in some cases, none- separately. Some letters have a similar shape with different positions of their dot(s). In addition, the jagged format of some letters and overlapping the letters in the words are other challenges. Based on extracting method of instances, traffic signs with real challenges, the proposed dataset includes six general challenges categories: weather conditions, lighting conditions, distance, background, color, and view angle.

As a step toward recognizing the introduced dataset, the single-stage deep learning technique of YOLOv3 with the Tesseract engine has been used to recognize cursive and multilingual languages. The CRAFT model was used to prepare this dataset based on deep learning techniques with 0.9705 precision and 0.9822 recall in scene text detection. In addition, YOLOv4 and YOLOv5 Algorithms have been compared with YOLOv3. The accuracy and F1-score values, evaluation criteria in YOLOv5, on the PESTD have been attained at 95.3% and 92.3%, respectively. The experiments depict the accuracy value, in YOLOv5 as 1.2% and 3.3% upper than YOLOv4 and YOLOv3, respectively. In addition, the F1-score criterion in YOLOv5 is 1.2% and 5.0% more than YOLOv4 and YOLOv3, respectively. Also, the calculation time of the YOLOv5 is 1.9 s and 7.2 s faster. As a future work, the authors plan to expand the database and add a large variety of traffic symbols. In addition, it will be great full to introduce new methods for recognizing scene images with higher accuracy.

Funding The author(s) received no financial support for this article’s research, authorship, and publication.

Data availability The datasets generated during the current study are available in the Persian-English-Scene-Text-Dataset (PESTD) repository, [Link].

Declarations

Conflict of interest The author(s) declared no potential conflicts of interest concerning this article’s research, authorship, and publication.

References

1. Ahmed SB, Naz S, Razzak MI, Yusof RB (2019) A novel dataset for English-Arabic scene text recognition (EASTR)-42 K and its evaluation using invariant feature extraction on detected extremal regions. IEEE Access
2. Baek Y, Lee B, Han D, Yun S, Lee H (2019) 'Character region awareness for text detection', in roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
3. Bochkovskiy A, Wang CY, Liao HYM (2020) 'Yolov4: Optimal speed and accuracy of object detection', arXiv preprint arXiv:2004.10934
4. Brunessaux S, Giroux P, Grilheres B, Manta M, Bodin M, Choukri K, Galibert O, Kahn J (2014) 'The Maurdor project: improving automatic processing of digital documents', 11th IAPR International Workshop on Document Analysis Systems (DAS), 349–354
5. Chernyshova Y, Emelianova E, Sheshkus A, Arlazarov VV (2021) 'MIDV- LAIT: A Challenging Dataset for Recognition of IDs with Perso-Arabic, Thai, and Indian Scripts', in International Conference on Document Analysis and Recognition, 258–272
6. Chowdhury MA, Deb K (2013) Extracting and Segmenting Container Name from Container Images. Int J Comput Appl 74:18–22
7. Chtourou I, Rouhou AC, Jaiem FK, Kanoun S (2015) 'ALTID: Arabic/Latin text images database for recognition research', in Document Analysis and Recognition (ICDAR), in 13th International Conference on, 836–840
8. Dvorin Y, Havosha UE (2009) 'Method and device for instant translation', Google Patents
9. Greenwood PM, Lenneman JK, Baldwin CL (2022) Advanced driver assistance systems (ADAS): Demographics, preferred sources of information, and accuracy of ADAS knowledge. Transport Res F: Traffic Psychol Behav 86:131–150
10. Gupta A, Vedaldi A, Zisserman A (2016) 'Synthetic data for text localisation in natural images', in Proceedings of the IEEE conference on computer vision and pattern recognition
11. 'International Phonetic Association and International Phonetic Association Staff and others, Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet', Cambridge University Press, 1999.
12. Jaderberg M, Simonyan K, Vedaldi A, Zisserman A (2014) 'Synthetic data and artificial neural networks for natural scene text recognition', in arXiv preprint arXiv:1406.2227
13. Karatzas D, Shafait F, Uchida S, Iwamura M, i Bigorda, LG, Mestre SR, Mas J, Mota DF, Almazan JA, De Las Heras LP (2013) 'ICDAR 2013 robust reading competition', in 12th International Conference on Document Analysis and Recognition, IEEE
14. Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar VR, Lu S, et al (2015) 'ICDAR 2015 competition on robust reading', in 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE
15. Kheirinejad S, Riaihi N, Azmi R (2020) 'Persian Text Based Traffic sign Detection with Convolutional Neural Network: A New Dataset', in 10th International Conference on Computer and Knowledge Engineering (ICCKE), IEEE
16. Lucas SM, Panaretos A, Sosa L, Tang A, Wong S, Young R (2003) 'ICDAR 2003 robust reading competitions', in Seventh International Conference on Document Analysis and Recognition, Proceedings, Springer
17. Maier D, Baden C, Stoltenberg D, De Vries-Kedem M, Waldherr A (2022) Machine translation vs. multilingual dictionaries assessing two strategies for the topic modeling of multilingual text collections. Commun Methods Meas 16(1):19–38
18. Mishra A, Alahari K, Jawahar C (2012) 'Top-down and bottom-up cues for scene text recognition', in IEEE Conference on Computer Vision and Pattern Recognition, IEEE
19. Mseddi WS, Sedrine MA, Attia R (2021) 'YOLOv5 Based Visual Localization for Autonomous Vehicles', in 29th European Signal Processing Conference (EUSIPCO), 746–750
20. Naiemi F, Ghods V, Khalesi H (2022) Scene text detection and recognition: a survey. Multimed Tools Appl 81:1–36
21. Phan TQ, Shivakumara P, Tian S, Tan CL (2013) 'Recognizing text with perspective distortion in natural scenes'
22. Powers DM (2020) 'Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation', arXiv preprint arXiv:2010.16061
23. Rashtehroudi AR, Shahbahrami S, Akoushideh A (2020) 'Iranian license plate recognition using deep learning', in International Conference on Machine Vision and Image Processing (MVIP)
24. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv preprint. https://doi.org/10.48550/ARXIV.1804.02767
25. Redmon J, Divvala S, Girshick R, Farhadi A (2016) 'You only look once: Unified, real-time object detection', in Proceedings of the IEEE conference on computer vision and pattern recognition

26. Schulz R, Talbot B, Lam O, Dayoub F, Corke P, Upcroft B, Wyeth G (2015) 'Robot navigation using human cues: A robot navigation system for symbolic goal-directed exploration', in International Conference on Robotics and Automation (ICRA)
27. Shetty AK, Saha I, Sanghvi RM, Save SA, Patel YJ (2021) 'A review: Object detection models', in 6th International Conference for Convergence in Technology (I2CT), 1–8
28. Tounsi M, Moalla I, Alimi AM, Lebouregois F (2015) 'Arabic characters recognition in natural scenes using sparse coding for feature representations', in 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE
29. Tounsi M, Moalla I, Alimi AM (2017) ARASTI: a database for Arabic scene text recognition. In 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), Nancy, France, pp 140–144. https://doi.org/10.1109/ASAR.2017.8067776
30. Tourani A, Soroori S, Shahbahrami A, Akoushideh A (2021) 'Iranis: A Large-scale Dataset of Iranian Vehicles License Plate Characters', in 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), pp. 1–5, https://doi.org/10.1109/IPRIA53572.2021.9483461
31. Tsai SS, Chen H, Chen D, Schroth G, Grzeszczuk R, Girod B (2011) 'Mobile visual search on printed documents using text and low bit-rate features', in 18th IEEE International Conference on Image Processing
32. Veit A, Matera T, Neumann L, Matas J, Belongie S (2016) 'Coco-text: Dataset and benchmark for text detection and recognition in natural images', in arXiv preprint arXiv:1601.07140
33. Wang K, Wei Z (2022) YOLO V4 with hybrid dilated convolution attention module for object detection in the aerial dataset. Int J Remote Sens 43(4):1323–1344
34. Wang K, Babenko B, Belongie S (2011) 'End-to-end scene text recognition', in International Conference on Computer Vision, IEEE
35. Wolf C, Jolion J (2006) Object count/area graphs for the evaluation of object detection and segmentation algorithms. IJDAR 8(4):280–296
36. Wu W, Liu H, Li L, Long Y, Wang X, Wang Z, Li J, Chang Y (2021) Application of local fully Convolutional Neural Network combined with YOLO v5 algorithm in small target detection of remote sensing image. PLoS One 16(10):e0259283
37. Yousfi S, Berrani S, Garcia C (2015) 'ALIF: A dataset for Arabic embedded text recognition in TV broadcast', in 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE
38. Zayene O, Hennebert J, Touj SM, Ingold R, Amara NEB (2015) 'A dataset for Arabic text detection, tracking and recognition in news videos-AcTiV', in 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE
39. Zhan F, Lu S, Xue C (2018) 'Verisimilar image synthesis for accurate detection and recognition of texts in scenes', in Proceedings of the European Conference on Computer Vision (ECCV)
40. Zhang C, Ding W, Peng G, Fu F, Wang W (2020) Street View Text Recognition With Deep Learning for Urban Scene Understanding in Intelligent Transportation Systems. IEEE Trans Intell Transp Syst 22:4727–4743
41. Zhu Y, Yao C, Bai X (2016) Scene text detection and recognition: Recent advances and future trends. Front Comput Sci 10(1):19–36

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.