SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

FATR: A Comprehensive Dataset and Evaluation Framework for Persian Text Recognition in Wild Images

Z. Raisi [*], *V. M. Nazarzehi Had, E. Sarani, R. Damani*

Electrical Engineering Department, Chabahar Maritime University, Chabahar, Iran.

Article Info

Article History:
Received 06 September 2024
Reviewed 11 Nvember 2024
Revised 27 December 2024
Accepted 30 December 2024

Keywords:
Persian scripts
Scene text recognition
Real-World datasets
Synthetic images
Deep learning
Farsi

[*] Corresponding Author's Email Address: *zobeir.raisi@cmu.ac.ir*

Abstract

Background and Objectives: Research on right-to-left scripts, particularly Persian text recognition in wild images, is limited due to lacking a comprehensive benchmark dataset. Applying state-of-the-art (SOTA) techniques on existing Latin or multilingual datasets often results in poor recognition performance for Persian scripts. This study aims to bridge this gap by introducing a comprehensive dataset for Persian text recognition and evaluating SOTA models on it.

Methods: We propose a Farsi (Persian) text recognition (FATR) dataset, which includes challenging images captured in various indoor and outdoor environments. Additionally, we introduce FATR-Synth, the largest synthetic Persian text dataset, containing over 200,000 cropped word images designed for pre-training scene text recognition models. We evaluate five SOTA deep learning-based scene text recognition models using standard word recognition accuracy (WRA) metrics on the proposed datasets. We compare the performance of these recent architectures qualitatively on challenging sample images of the FATR dataset.

Results: Our experiments demonstrate that SOTA recognition models' performance declines significantly when tested on the FATR dataset. However, when trained on synthetic and real-world Persian text datasets, these models demonstrate improved performance on Persian scripts.

Conclusion: Introducing the FATR dataset enhances the resources available for Persian text recognition, improving model performance. The proposed dataset, trained models, and code is available at https://github.com/zobeirraisi/FATDR.

Introduction

Text is a crucial source of visual information in our daily lives. It can be found everywhere, from documents and images to street signs, billboards, house numbers, and license plates. These texts provide vital details about location and identity and have various applications in real life [1]-[4]. Identifying text from input images involves two primary steps: first, accurately localizing the text instance (scene text detection), and second, converting the detected regions into word or character strings (scene text recognition).

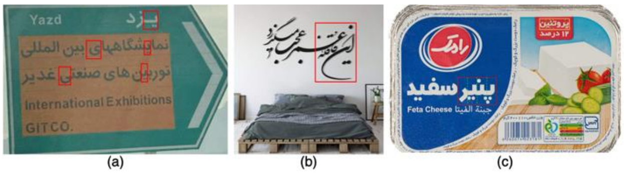


Fig. 1: The challenges of Persian scripts in the wild images. (a) the same character with identical font, as seen inside of the red box, can appear in different shapes according to its position in the word instances, (b) a high degree of overlap in characters, and (c) The first three characters of the word ”ی ” are distinct characters, distinguished by the arrangement of small dots known as ”Noghteh”.

Doi: 10.22061/jecei.2024.11256.784

Table 1: Persian characters with similar body shapes

ج چ ح خ	ب پ ت ث	آا
س ش	رزژ	دذ
ع غ	طظ	ص ض
ک گ	ق	ف
ن	م	ل
ی	و	ه

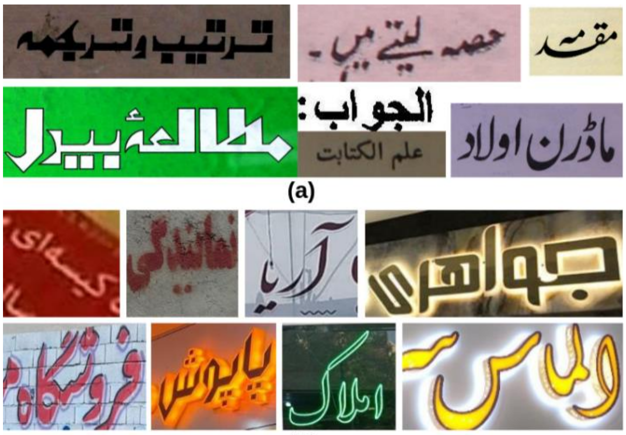
Detecting and recognizing text in images with diverse characteristics such as color, font, orientation, language, and scene complexity is challenging. Traditional classical machine learning methods [5], [6] often struggle with complex scenarios. In contrast, recent deep learning-based approaches [7]-[15] have shown promising results in detecting and recognizing text even under hostile conditions. However, the majority of recent scene text detection and recognition has been conducted on Latin scripts, resulting in the development of multiple benchmark datasets for this purpose. This focus on Latin scripts has created a significant gap in text detection and recognition for non-Latin languages that use right-to-left scripts, such as Persian, Arabic, and Urdu. These languages have unique features that distinguish them from Latin writing systems, highlighting the importance of addressing this challenge with more attention.

Persian text recognition in the wild is a more challenging task due to its unique characteristics that differ significantly from Latin scripts. The complexity of this challenge is illustrated in Fig. 1 As seen, these challenges are connected letters from different positions (front, back, or side), diacritical marks, and the same character appearing differently in different positions of word instances (as shown in Fig. 1(a)), overlapping characters (shown in Fig. 1(b)). In addition, the Persian script is full of another specific challenge different from Latin text instances and that is the appearance of identical shapes characters with a different number of placement of dots (as shown in Fig. 1(c)) that causes problems in recognizing of these characters, which are illustrated in Table 1. These challenges independently pose a significant obstacle to current SOTA text detection and recognition (TDR) methods, which are mainly designed for Latin scripts. Persian letters are either horizontally or vertically oriented, with horizontal letters playing a crucial role in connectivity.

In contrast to Arabic and Urdu scripts, publicly available datasets for Persian scene text recognition are limited (See Section 2.2 for more details). While Persian scripts are similar to Arabic and Urdu, using existing Arabic or Urdu benchmark datasets may lead to poor performances. For instance, as seen in Table 2, the most recent Urdu scene text recognition model [16] that includes all classes of Persian alphabets in Table 3, still falls short of expected levels of word recognition accuracy.

Table 2: Comparing the word recognition accuracy (see Section 4.2.1) performance of the Urdu language model proposed in [16] on Urdu and Persian datasets. We used the test set of the cropped word images of our proposed dataset. Some sample images of both languages are provided in Fig. 2

Model	Urdu Dataset	Persian Dataset
Urdu-Large	92.97	**38.37**



(a)

(b)

Fig. 2: Comparison of (a) Urdu script and (b) Farsi script, where the model in Table 2 (Urdu-Large [16]) successfully recognized all the images in (a) while failing on images in (b).

As seen from Table 2 the UTR-Net with a WRA of 92.97% declines significantly ~50% on Persian scripts. Fig. 2 demonstrates some sample images of both languages with similar characters but different styles and fonts that the model in [16] successfully recognized the images in Fig. 2(a) while failing or missing some characters in Fig. 2(b) . Therefore, introducing or preparing a unique dataset for the Persian language is essential. Furthermore, the earlier proposed dataset for Persian scripts focused on only offering a synthetic dataset as in [17], focusing on single task detection as in [18], [19] or recognition or a specific kind of text instances like documents as in [20], [21], [63].

To address the mentioned problems, we introduce a new dataset for detecting and recognizing Persian text in real-world situations. The proposed FATR dataset is designed to be comprehensive and a good benchmark for measuring the robustness and generalizability performance of current and future models. To prepare this dataset, we captured a diverse collection of in-the-wild images tailored to the unique features of Persian script. We also built a large-scale synthetic Persian text dataset that can be used for training and evaluating Persian scene text recognition models. By addressing the challenges of real-life scenarios, our study advances the field of text recognition, bridging the gap between Table 3.

Table 3: The Persian characters with their appearance in scripts

Persian Letter	Symbol	beginning	middle	end
همزه	ء أ	ئـ	ـئـ	ـئ ـأ ـؤ
الف	ا	آ		
بِ	ب	بـ	ـبـ	ـب
پِ	پ	پـ	ـپـ	ـپ
تِ	ت	تـ	ـتـ	ـت
ثِ	ث	ثـ	ـثـ	ـث
جیم	ج	جـ	ـجـ	ـج
چِ	چ	چـ	ـچـ	ـچ
حِ	ح	حـ	ـحـ	ـح
خِ	خ	خـ	ـخـ	ـخ
دال	د			ـد
ذال	ذ			ـذ
رِ	ر			ـر
زِ	ز			ـز
ژِ	ژ			ـژ
سین	س	سـ	ـسـ	س
شین	ش	شـ	ـشـ	ش
صاد	ص	صـ	ـصـ	ص
ضاد	ض	ضـ	ـضـ	ض
طا	ط	طـ	ـطـ	ـط
ظا	ظ	ظـ	ـظـ	ـظ
عین	ع	عـ	ـعـ	ـع
غین	غ	غـ	ـغـ	ـغ
فِ	ف	فـ	ـفـ	ـف
قاف	ق	قـ	ـقـ	ـق
کاف	ک	کـ	ـکـ	ـک
گاف	گ	گـ	ـگـ	ـگ
لام	ل	لـ	ـلـ	ـل
میم	م	مـ	ـمـ	ـم
نون	ن	نـ	ـنـ	ـن
واو	و			ـو
هِ	ه	هـ	ـهـ	ـه
یِ	ی	یـ	ـیـ	ـی

The main contributions are summarized as follows:

1. We propose a Persian text recognition dataset. To the best of our knowledge, this is the first publicly available dataset that contains various text instances captured in wild images from different environments, considering all the challenges in Latin benchmark datasets. This dataset can be used as a benchmark for future research.
2. We also present a large-scale synthetic dataset for Persian scent text recognition of about 200K cropped word images.
3. We review the past and recent advancements in scene text recognition for Latin and non-Latin scripts.
4. We train six well-known SOTA scene text recognition model, and evaluate, compare, and analyze quantitatively and qualitatively their performances on the proposed FATR dataset.

Related Work

A. Scene Text Recognition

In scene text recognition, the main objective is identifying the characters or words present within the detected text regions in the given input images. This task is more complex than recognizing printed scanned documents because real-world images pose various challenges, such as low resolution, extreme lighting, diverse fonts, orientations, languages, and lexicons compared to the clean background scanned or printed documents. To address these difficulties, researchers have proposed different methods based on both classical machine learning techniques such as [22]-[24], and deep learning techniques such as [10]-[13], [25].

Classical machine learning-based methods [22], [26], [27] typically use features like HOG [28] or SIFT [29] in combination with classifiers such as SVM [30]. These methods either adopt a bottom-up approach, where classified characters are linked into words, or a top-down approaches that directly recognize entire words from the image [31]. However, these methods often struggle to recognize new words that are not part of the training dataset, and they have limited capabilities in representing features that are essential for real-world scenarios. Additionally, classical methods are often unable to recognize input word images that are multi-oriented or curved, which are common in wild images.

On the other hand, recent scene text recognition methods utilize deep learning architectures to address the challenges of complex real-world scenarios. Inspired by speech recognition, many recent methods model scene text as a sequence of characters [10]-[12], which are called sequence-based methods. These methods leverage techniques like Connectionist Temporal Classification (CTC) [32] to predict character sequences. However, these methods are designed for 1-dimensional (1D) sequences, and converting 2D image features to 1D leads to information loss, hindering the recognition of irregular text. To address this, researchers proposed a 2D-CTC [33] technique that directly operates on 2D probability distributions, achieving better recognition accuracy.

The attention mechanism initially used for machine translation has also been adopted for scene text recognition [11], [34]. Attention allows the model to focus on specific image regions during decoding, enhancing the recognition of irregular text. Different attention-based frameworks have been proposed, ranging from basic 1D-attention models to more complex methods that employ rectification or character-aware techniques to handle various text distortions. However, some methods have difficulty recognizing images with complex backgrounds or high computational costs. With the advancement of the transformer architecture [35], many recent scene text recognition models [36], [37] have utilized the transformer in their pipeline and achieved SOTA performance in several benchmark datasets with complex and challenging word images.

B. *Datasets*

Latin Scripts: The scene text recognition benchmarks can be categorized into two general categories: regular text datasets, including ICDAR13 [38], III5k [39], and SVT [23], which contain primarily horizontal text instances, and irregular text datasets, including ICDAR15 [40], CUT80 [41], SVT-P [42], and COCO-Text [43], which contain challenging multi-oriented and curved text instances.

Researchers also pre-trained their models on synthetic images to achieve a more general and higher accuracy performance. SynthText (ST) [44] and MJSynth (MJ) [45] synthetic datasets are two datasets that have been used extensively for the pre-training purposes of scene text detection and recognition algorithms.

Multi-Lingual Scripts: Researchers used several multi-lingual text datasets to measure the performance of their models. ICDAR17-MLT [46] and ICDAR19-MLT [47] are two examples of multilingual datasets that contain the following languages: Arabic, Latin, Chinese, Japanese, Korean, Bangla, and Hindi. There are also some other datasets [48]-[51] that have scripts, mostly in English and Chinese, that are designed for the specific purposes of text recognition.

Right-to-Left Scripts: Arabic, Urdu, and Persian are three languages that use similar letter scripts but are distinct when spoken. Different from left-to-right languages such as Latin and Chinese, finding publicly available benchmark datasets specifically designed for these languages can be challenging. However, numerous conventional and modern techniques have been developed on private datasets for these languages. In this case, Arabic and Urdu are better suited than Persian. For instance, ICDAR17-MLT [46] and ICDAR19-MLT [47] are two publicly available benchmarks that contain Arabic scripts and can be utilized for training and evaluating Arabic text recognition. ARASTEC [52] and ARASTI [53] are two real-world datasets used for Arabic character and word recognition in natural images, respectively. However, these datasets are not available to the public. For Urdu text, IIITH [54] and UPTI [55] are two well-known Urdu script datasets that include real-world and synthetic text instances. Recently, a work by Rahman et al. [16] introduced the UTRSet-Real and UTRSet-Synth datasets, which are publicly available.

Regarding Persian, the only real-word dataset currently is PESTD [18]. This dataset is unique in that it is a Persian English dataset with images captured in the wild. However, it is only designed for traffic sign detection and mainly has images of road traffic signs. Moreover, it has yet to consider the recognition task, which is the most challenging aspect. Furthermore, the dataset is private for testing. A publicly available dataset contains synthetic images, namely ITDR-Synth [17], designed for both detection and recognition. This dataset contains 6,100 and 40,220 images for detection and recognition, respectively.

Farsi Text Recognition (FATR) Dataset

In this section, we present our comprehensive Persian language dataset tailored specifically for investigating and analyzing Persian text recognition challenges. For this purpose, a synthetic dataset for the recognition task that contains text instances of captured images of both indoor and outdoor environments. Table 4 shows more details of the proposed FATR dataset. The indoor images contain images of dense and small text instances taken from indoor store signs and products. All the outdoor images consist of a wide variety of challenging cases of urban landscapes, such as storefronts, street signs, wall signs, and traffic signs.

Table 4: The proposed Farsi (Persian) text dataset

Text Category		Recognition	
		Train	Test
Indoor Text	Product	648	158
	Lobbies	2080	748
Outdoor Text	Storefronts	7796	1951
	Street signs	261	81
	Graffiti	346	92
	Traffic signs	1804	501
All		12935	3529

Fig. 3 shows different sample images of FATR that are taken with different camera phones.

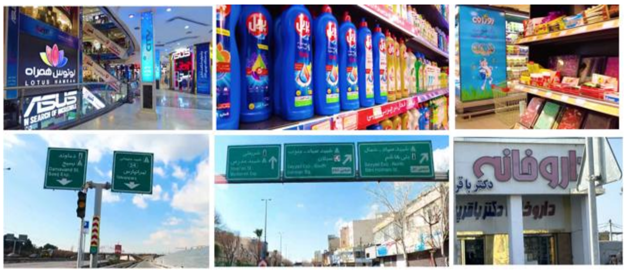


Fig. 3: Sample real-world images of the proposed FATR dataset.

A. *Real-World Text Recognition*

We converted the quadrilateral boxes into rectangular boxes and cropped them to prepare the recognition dataset. The recognition dataset consists of 12935-word images cropped for training and 3529 for testing. In Fig. 4, you can see some sample images and their corresponding strings.

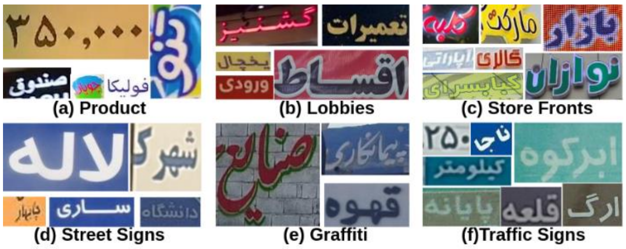


Fig. 4: Example of real-world cropped word patch images of the proposed FATR dataset used for training and evaluation of recognition model taken from different indoor and outdoor places.

We consider various challenges when preparing the recognition images of FATR datasets. Fig. 5 illustrates some sample images of these challenges including partially occluded text, rotated text, illumination variation, low resolution, English text, image blurriness, complex background, difficult fonts, and special characters (*e.g.*, /,() -,:, @,#,…).

We provide a probability distribution of the cropped word images in the FATR dataset, focusing on image height, width, and character length. Fig. 6 illustrates these computations. As shown in Fig. 6(a) and Fig. 6(b), the FATR dataset contains word images of varying resolutions, with a minimum width of 4 pixels a minimum height of 5 pixels, and a maximum width of 4391 pixels, and a maximum height of 2505 pixels.

As seen in Fig. 6(c), the length of the word instances ranges from 1 character to 22 characters, with an average length of approximately 5 characters per word. It is worth mentioning that the dataset contains a total of 5795 unique word instances.

B. *Synthetic Persian Dataset*

The SOTA scent text techniques are trained on a combination of large synthetic cropped word images of SynthText [44] and MJ-Synth [45], and they achieved good performances on real-world benchmark datasets. In this paper, we also introduce FATR-Synth, a synthetic dataset of ~200K word images of Persian scripts, and the real-world images of FATR for training the SOTA scene text recognition models. Inherited from [16], these images are created using different text attributes like font, size, and color and include over 200 Persian fonts. It addresses the scarcity of Persian words and numerals in existing datasets by incorporating sufficient samples and provides a vocabulary of 200,000 words and ~50000 unique Farsi words collected from the Internet with an average length of 8 characters for synthetic text generation. The produced dataset is also publicly available for further research. Fig. 7 illustrates some examples of the prepared synthetic dataset.

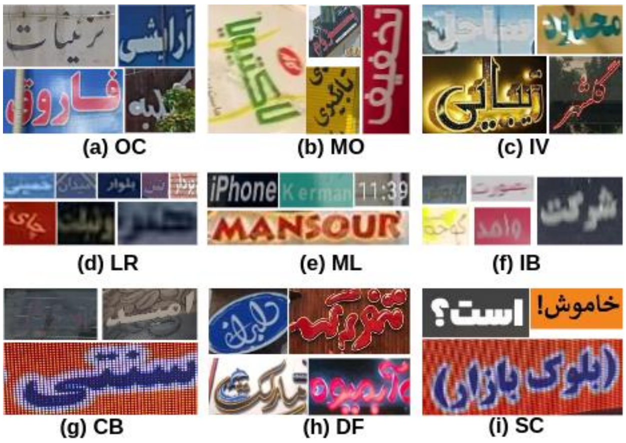


Fig. 5: Challenges in real-world images from the FATR dataset. The identified challenges include OC (occluded text), MO (multi-oriented), IV (illumination variation), LR (low resolution), ML (multi-language), IB (image blurriness), CB (complex background), DF (difficult fonts), and SC (special characters).

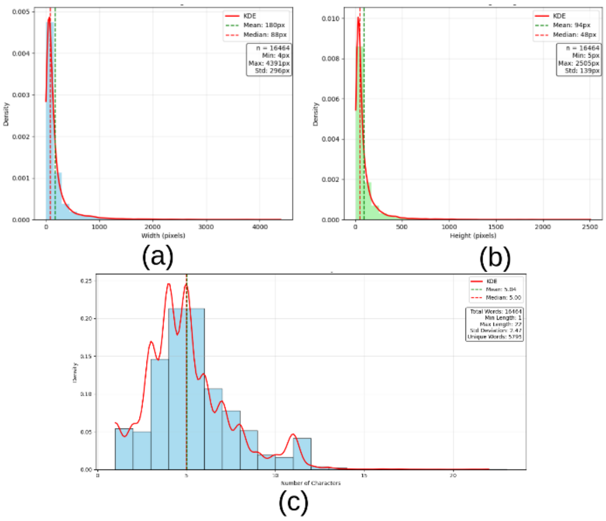


Fig. 6: Probability distribution of (a) word width in pixels, (b) word height in pixels, and (c) the number of characters in a word image computed from the FATR dataset. Best viewed when zoomed.

Fig. 7: Sample synthetic word images of the proposed FATRSynth dataset. This dataset is only used as training. Each cropped word image mainly contains one-word instances.

Experimental Results

This section presents our comprehensive evaluation for investigating some selected SOTA text recognition models including CRNN [10], STAR-Net [12], ROSETTA [13], CLOVA [14], and UTR-Net [16] on the proposed FATR dataset. Table 4 shows the number of test images used for the evaluation of these models.

A. Implementation Details

All models used in this paper are trained and tested on a machine equipped with an NVIDIA GPU with plate number RTX-3090. To ensure a fair comparison, we train all the models in comparisons on a similar dataset. For the recognition task, we follow the same settings described in [14] to train the models. We use a combination of FATR-Synth and FATR real-world images for training. To generate the images of FATR-Synth, we follow the settings provided in [16][1].

In this paper, we only focus on evaluating Persian characters. Therefore, we only consider Persian text instances during inference and ignore English text. We evaluate the selected recognition models using the Word Recognition Accuracy (WRA) and Normalized Edit Distance (NED) evaluation metrics. We train all recognition models, except UTRNeT [16], for 200,000 iterations, while the UTRNeT model is trained for 50 epochs. We use the following 35 characters for training all the evaluation models: ا, آ, ب, پ, ت, ث, ج, چ, ح, خ, د ,
ذ, ر, ز, ژ, س, ش, ص, ض, ط, ظ, ع, غ, ف, ق, ک, گ, ل ,
م, ن, و, ه, ء, ی, ئ.

B. Evaluation Metrics

For the recognition task, given a set of cropped word images, we use two metrics for the evaluation of recognition models: Word Recognition Accuracy (WRA) and Normalized Edit Distance (NED). WRA is mainly used to evaluate the accuracy of scene text recognition schemes [10], [12]-[14]., which can be calculated as:

$$\text{WRA} = \frac{\#\ \text{accurately Recognized Words}}{\mathit{All\ the\ word\ instances}} \times 100 \qquad (1)$$

The NED metric is defined as follows [47]:

$$Norm = 1 - \frac{1}{N}\sum_{i=1}^{N} D(w_i, w_i') / \max(w_i, w_i') \qquad (2)$$

where Levenshtein Distance is shown by $D(:)$ [56]. w_i and w_i' are ground truths corresponding to the text regions and predicted word strings, respectively.

C. Quantitative Results

We conducted several experiments on the recognition using pre-training models. Persian scripts' characters are fundamentally different from those of left-to-right languages, and using a SOTA right-to-left language model would result in a significantly low WRA margin (See Table 2 in Section 1). Therefore, we selected some of the best and most well-known recognition models [10], [12]-[14], [16], trained them in a similar setting, and evaluated them on the FATR dataset. Table 5 shows the quantitative results of our experiments. We first trained these models on synthetic images in the ITDR-Synth dataset, which resulted in poor performance. However, when we used our proposed FATR-Synth dataset, the models' WRA performance improved significantly. We further enhanced the models' performance by combining natural and synthetic images of the proposed FATR dataset. Among these models, CLOVA [14] achieved the best performance regarding both WRA and edit distance (ED) for all our experiments. Our quantitative results confirm that training the models on the proposed FATR dataset can significantly improve the recognition performance compared to the only publicly available synthetic Persian dataset [17].

Table 5: Experimental Results of the select scene text recognition models [10], [12]-[14], [16] on the proposed FATR dataset. The results trained on the synthetic images of ITDR-Synth proposed in [17] and our proposed FATR-Synth are shown with blue and red colors, respectively. All the model results trained on the combination of our proposed synthetic and real-world images of FATR are shown in black color. The WRA and NED denote the word recognition accuracy and the normalized edit distance.

Model	Trained Dataset	WRA	NED
CRNN [10]	IDTR-Synth	22.28	0.45
	FATR-Synth	45.65	0.75
	FATR-Synth+FATR-Real	53.04	0.78
ROSETTA [13]	IDTR-Synth	19.26	0.44
	FATR-Synth	36.84	0.72
	FATR-Synth+FATR-Real	65.7	0.85
STARNET [12]	IDTR-Synth	27.17	0.48
	FATR-Synth	64.66	0.85
	FATR-Synth+FATR-Real	68.74	0.86
CLOVA [14]	IDTR-Synth	27.68	0.50
	FATR-Synth	64.94	0.85
	FATR-Synth+FATR-Real	69.24	0.87
UTRNet [16]	IDTR-Synth	24.72	0.49
	FATR-Synth	52.98	0.77
	FATR-Synth+FATR-Real	66.93	0.86

D. Qualitative Results

We tested the models listed in Table 5 to demonstrate their performance on real-world images. To that effect, we evaluated the qualitative results on various cropped word images from the FATR dataset, as presented in Fig. 8. The output strings of Fig. 8(a)-(c) demonstrate that the chosen models can accurately recognize regular text

[1] https://github.com/abdur75648/urdu-synth/

instances with horizontal or near-horizontal orientation. However, some selected models failed or produced the wrong output for one or two characters when evaluated on challenging examples, such as rotated text or text with complicated font styles, as shown in Fig. 8(d)-(g). Ultimately, all the models we evaluated produced false recognition when the text was vertically oriented, partially occluded, or used a complex font in adverse situations.

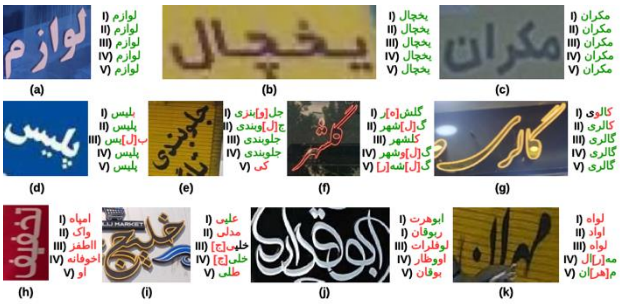


Fig. 8: Qualitative results among the selected recognition models [10], [12]-[14], [16] on some images of the FATR dataset. Each output strings stand for the following models: I) CRNN [10], II) ROSETTA [13], III) STAR-Net [12], IV) CLOVA [14], and V) UTRNet [16]. These models are trained on the combination of real and synthetic images of the FATR dataset. The green and red colors denote the accurate and inaccurate characters predicted by the models.

E. Discussion and Future Work

Persian text recognition in the wild presents significant challenges, one of which is the time-consuming and costly annotation process. Recent advancements in artificial intelligence, such as DALLE [57], ChatGPT [58], and Gemini [59], have made it possible to generate images using text prompts, providing a way to address this issue. Another approach to tackle the annotation problem is automating the process, which can be achieved using the latest model [60]. However, detecting and recognizing Latin text, as well as text images taken from the wild, still pose various challenges, such as orientation, occlusion, and degradation in image quality. These challenges remain unsolved problems in the computer vision community. To overcome these challenges, techniques like augmentation, compositionality [61], and masked autoencoder transformers [62] combined with AI modules can be used to assist the models. As future work, we also aim to design a deep learning-based architecture by utilizing the above-mentioned advancement in computer vision and natural language processing to tackle the shortcomings of the current state-of-the-art models and capture and recognize challenging word images from wild images.

Conclusions

This paper highlights a critical research gap in right-to-left text recognition in wild images, specifically for Persian scripts. We have introduced a comprehensive Persian text recognition dataset to address this issue, which provides real-world and synthetic images to evaluate SOTA text recognition models. We have evaluated several scene text recognition models on the proposed FATR dataset. Our experimental results have shown that the current multilingual text dataset can still perform well on Persian scripts. For Persian scent text recognition, a specialized dataset is essential to accurately recognize text instances in wild images.

Acknowledgment

The authors express their acknowledgement to the JECEI reviewers, editors, and editorial board for their constructive feedback, valuable suggestions, and professional support, which improved the quality of our paper.

Author Contributions

Z. Raisi collected the data, implemented the code, carried out the analysis, and wrote the paper. The other authors also equally collected the data, edited the paper and interpreted the results.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication or falsification, double publication and, or submission, and redundancy, have been completely witnessed by the authors.

Abbreviations

FATR	Farsi (Persian) text recognition
SOTA	State-of-The-Art
WRA	Word Recognition Accuracy
NED	Normalized Edit Distance
CTC	Connectionist Temporal Classification

References

[1] Y. Zhu, C. Yao, X. Bai, “Scene text detection and recognition: Recent advances and future trends,” Front. Comput. Sci., 10(1): 19-36, 2016.

[2] H. Lin, P. Yang, F. Zhang, “Review of scene text detection and recognition,” Arch. Comput. Methods Eng., 27: 433-454, 2020.

[3] Z. Raisi, M. A. Naiel, P. Fieguth, S. Wardell, J. Zelek, “Text detection and recognition in the wild: A review,” arXiv preprint arXiv:2006.04305, 2020.

[4] Z. Raisi, J. Zelek, “Text detection and recognition for robot localization,” J. Electr. Comput. Eng. Innov., 12(1): 163-174, 2024.

[5] K. Wang, B. Babenko, S. Belongie, “End-to-end scene text recognition,” in Proc. 2011 International Conference on Computer Vision: 1457-1464, 2011.

[6] A. Bissacco, M. Cummins, Y. Netzer, H. Neven, “PhotoOCR: Reading text in uncontrolled conditions,” in Proc. 2013 IEEE International Conference on Computer Vision: 785-792, 2013.

[7] Z. Raisi, V. M. Nazarzehi, “A transformer-based approach with contextual position encoding for robust persian text recognition in the wild,” J. AI Data Min., 12(3): 455-464, 2024.

[8] Z. Raisi, G. Younes, J. Zelek, “Arbitrary shape text detection using transformers,” in Proc. 2022 26th International Conference on Pattern Recognition (ICPR): 3238-3245, 2022.

[9] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman., “Deep structured output learning for unconstrained text recognition,” arXiv:1412.5903v5, 2015.

[10] B. Shi, X. Bai, C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” IEEE Trans. Pattern Anal. Mach. Intell., 39(11): 2298-2304, 2016.

[11] B. Shi, X. Wang, P. Lyu, C. Yao, X. Bai, “Robust scene text recognition with automatic rectification,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 4168- 4176, 2016.

[12] W. Liu, C. Chen, K. Y. K. Wong, Z. Su, J. Han, “STARNet: A spatial attention residue network for scene text recognition,” in Proc. British Machine Vision Conference (BMVC): 43.1-43.13, 2016.

[13] F. Borisyuk, A. Gordo, V. Sivakumar, “Rosetta: Large scale system for text detection and recognition in images,” in Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining: 71-79, 2018.

[14] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, H. Lee, “What is wrong with scene text recognition model comparisons? Dataset and model analysis,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV): 4715-4723, 2019.

[15] C. Ma, L. Sun, J. Wang, Q. Huo, “Dq-detr: Dynamic queries enhanced detection transformer for arbitrary shape text detection,” in Proc. International Conference on Document Analysis and Recognition: 243-260, 2023.

[16] A. Rahman, A. Ghosh, C. Arora, “Utrnet: Highresolution urdu text recognition in printed documents,” in Proc. International Conference on Document Analysis and Recognition: 305-324, 2023.

[17] F. Alimoradi, F. Rahmani, L. Rabiei, M. Khansari, M. Mazoochi, “Synthesizing an image dataset for text detection and recognition in images,” J. Inf. Commun. Technol., 53(53): 78, 2023 [In Farsi].

[18] A. Rashtehroudi, A. Ranjkesh, A. Shahbahrami, "PESTD: a large-scale Persian-English scene text dataset," Multimedia Tools Appl., 82: 34793-34808, 2023.

[19] S. Kheirinejad, N. Riaihi, R. Azmi, “Persian text-based traffic sign detection with convolutional neural network: A new dataset,” in Proc. 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE): 060- 064, 2020.

[20] M. Rahmati, M. Fateh, M. Rezvani, A. Tajary, V. Abolghasemi, “Printed persian ocr system using deep learning,” IET Image Process., 14(15): 3920-3931, 2020.

[21] A. Fateh, M. Rezvani, A. Tajary, M. Fateh, “Persian printed text line detection based on font size,” Multimedia Tools Appl., 82(2): 2393-2418, 2023.

[22] T. E. De Campos, B. R. Babu, M. Varma, et al., “Character recognition in natural images,” in Proc. Fourth International Conference on Computer Vision Theory and Applications (VISAPP), 7: 273-280, 2009.

[23] K. Wang, S. Belongie, “Word spotting in the wild,” in Proc. European Conference on Computer Vision: 591-604, 2010.

[24] L. Neumann, J. Matas, “Real-time scene text localization and recognition,” in Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition: 3538-3545, 2012.

[25] F. Zhan, S. Lu, “Esir: End-to-end scene text recognition via iterative image rectification,” in Proc. 2019 IEEE Conference on Computer Vision and Pattern Recognition: 2059-2068, 2019.

[26] M. Sawaki, H. Murase, N. Hagita, “Automatic acquisition of context-based images templates for degraded character recognition in scene images,” in Proc. 15th International Conference on Pattern Recognition (ICPR), 4: 15-18, 2000.

[27] Y. F. Pan, X. Hou, C. L. Liu, “Text localization in natural scene images based on conditional random field,” in Proc. 2009 10th International Conference on Document Analysis and Recognition: 6-10, 2009.

[28] N. Dalal, B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1: 886-893, 2005.

[29] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. of Comp. Vision, 60(2): 91-110, 2004.

[30] J. A. Suykens, J. Vandewalle, “Least squares support vector machine classifiers,” Neural Process. Lett., 9(3): 293-300, 1999.

[31] J. Almazan, A. Gordo, A. Forn´ es, E. Valveny, “Word´ spotting and recognition with embedded attributes,” IEEE Trans. Pattern Anal. Mach. Intell., 36(12): 2552-2566, 2014.

[32] A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in Proc. 23rd International Conference on Machine Learning: 369-376, 2006.

[33] Z. Wan, F. Xie, Y. Liu, X. Bai, C. Yao, “2D-CTC for scene text recognition,” arXiv:1907.09705v1, 2019.

[34] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, X. Bai, “Aster: An attentional scene text recognizer with flexible rectification,” IEEE Trans. Pattern Anal. Mach. Intell., 41(9): 2035-2048, 2018.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, “Attention is all you need,” in Proc. 31st Conference on Neural Information Processing Systems (NIPS 2017): 5998-6008, 2017.

[36] Z. Raisi, M. A. Naiel, G. Younes, S. Wardell, J. Zelek, “2lspe: 2d learnable sinusoidal positional encoding using transformer for scene text recognition,” in Proc. 2021 18th Conference on Robots and Vision (CRV): 119-126, 2021.

[37] Z. Qiao, Z. Ji, Y. Yuan, J. Bai, “Decoupling visual semantic features learning with dual masked autoencoder for self-supervised scene text recognition,” in Proc. International Conference on Document Analysis and Recognition: 261-279, 2023.

[38] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, L. P. De Las Heras, “ICDAR 2013 robust reading competition,” in Proc. 2013 12th International Conference on Document Analysis and Recognition: 1484-1493, 2013.

[39] A. Mishra, K. Alahari, C. V. Jawahar, “Scene text recognition using higher order language priors,” in Proc. BMVC, 2012.

[40] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, et al., “ICDAR 2015 competition on robust reading,” in Proc. 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015.

[41] A. Risnumawan, P. Shivakumara, C. S. Chan, C. L. Tan, “A robust arbitrary text detection system for natural scene images,” Expert Syst. with Appl., 41(18): 8027- 8048, 2014.

[42] T. Quy Phan, P. Shivakumara, S. Tian, C. Lim Tan, “Recognizing text with perspective distortion in natural scenes,” in Proc. IEEE International Conference on Computer Vision (ICCV): 569-576, 2013.

[43] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C. L. Zitnick, “Microsoft coco: Com-´ mon objects in context,” in Proc. 13th European Conference on Computer Vision: 740-755, 2014.

[44] A. Gupta, A. Vedaldi, A. Zisserman, “Synthetic data for text localisation in natural images,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 2315-2324, 2016.

[45] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, “Synthetic data and artificial neural networks for natural scene text recognition,” arXiv preprint arXiv:1406.2227, 2014.

[46] M. Iwamura, N. Morimoto, K. Tainaka, D. Bazazian, L. Gomez, D. Karatzas, “ICDAR2017 robust reading challenge on omnidirectional video,” in Proc. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 1: 1448-1453, 2017.

[47] Y. Sun, Z. Ni, C. K. Chng, Y. Liu, C. Luo, C. C. Ng, J. Han, E. Ding, J. Liu, D. Karatzas, et al., “ICDAR 2019 competition on large-scale street view text with partial labeling– RRC-LSVT,” 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019.

[48] W. Wu, Y. Zhao, Z. Li, J. Li, M. Z. Shou, U. Pal, D. Karatzas, X. Bai, “Icdar 2023 competition on video text reading for dense and small text,” in Proc. International Conference on Document Analysis and Recognition: 405–419, 2023.

[49] R. Zhang, Y. Zhou, Q. Jiang, Q. Song, N. Li, K. Zhou, L. Wang, D. Wang, M. Liao, M. Yang, et al., “ICDAR 2019 robust reading challenge on reading Chinese text on signboard,” in Proc. 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019.

[50] C. K. Chng, Y. Liu, Y. Sun, C. C. Ng, C. Luo, Z. Ni, C. Fang, S. Zhang, J. Han, E. Ding, J. Liu, D. Karatzas, C. Seng Chan, L. Jin, “Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art,” in Proc. 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019.

[51] Z. Wan, J. Zhang, L. Zhang, J. Luo, C. Yao, “On vocabulary reliance in scene text recognition,” in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 11425-11434, 2020.

[52] M. Tounsi, I. Moalla, A. M. Alimi, F. Lebouregois, “Arabic characters recognition in natural scenes using sparse coding for feature representations,” in Proc. 2015 13th International Conference on Document Analysis and Recognition (ICDAR): 1036-1040, 2015.

[53] M. Tounsi, I. Moalla, A. M. Alimi, “Arasti: A database for arabic scene text recognition,” in Proc. 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR): 140-144, 2017.

[54] M. Jain, M. Mathew, C. Jawahar, “Unconstrained ocr for urdu using deep cnn-rnn hybrid networks,” in Proc. 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR): 747- 752, 2017.

[55] N. Sabbour, F. Shafait, “A segmentation-free approach to arabic and urdu ocr,” in Proc. Document recognition and retrieval XX, 8658: 215-226, 2013.

[56] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” in Soviet physics doklady, 10: 707-710, 1966.

[57] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, “Hierarchical text-conditional image generation with clip latents,” arXiv preprint arXiv:2204.06125, 1(2): 3, 2022.

[58] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., “Gpt-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.

[59] G. Team, R. Anil, S. Borgeaud, Y. Wu, J. B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., “Gemini: a family of highly capable multimodal models,” arXiv preprint arXiv:2312.11805, 2023.

[60] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, et al., “Segment anything,” in Proc. the IEEE/CVF International Conference on Computer Vision: 4015- 4026, 2023.

[61] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, A. Yuille, “Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion,” arXiv preprint arXiv:2006.15538, 2020.

[62] Z. Raisi, J. Zelek, “Occluded text detection and recognition in the wild,” in Proc. 2022 19th Conference on Robots and Vision (CRV): 140-150, 2022.

[63] A. Faraji, M. Saeed, H. Nezamabadi-pour, "Introducing a database for Farsi document image understanding and segmentation," J. Mach. Vision Image Process., 10(2): 31-46, 2023 [In Persian].

Biographies

Zobeir Raisi received his Ph.D. degree in 2022 from the Vision Image Processing Lab (VIPLab) at the Systems Design Engineering Department, University of Waterloo, Waterloo, Ontario, Canada. Currently, he is an Assistant Professor in the Department of Electrical Engineering at Chabahar Maritime University, Iran. His research interests include computer vision, artificial intelligence, and robotics.

- Email: zobei.raisi@cmu.ac.ir
- ORCID: 0000-0002-1591-4492
- Web of Science Researcher ID: GLV-1410-2022
- Scopus Author ID: 54897975500
- Homepage: https://www.cmu.ac.ir/staff/zraisi

Valimohammad Nazarzehi Had received his Ph.D. degree in 2016 from the University of New South Wales, Australia. Currently, he is an Assistant Professor in the department of Electrical Engineering, Chabahar Maritime University, Iran. His research interests include decentralized control, marine control systems, and control of mobile robots, robotics, and image processing.

- Email: v.nazarzehi@cmu.ac.ir
- ORCID: 0000-0003-3261-6320
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://www.cmu.ac.ir/staff/vnazarzehi

Esmaeil Sarani is a Ph.D. graduate in Electrical Power Engineering from the University of Tehran and currently serves as an Assistant Professor in the Department of Electrical Engineering at the Chabahar Maritime University, Iran. His research interests include the design of electrical machines, fault detection in electrical machines, renewable energy development with a focus on wave energy harvesting, and the application of artificial intelligence in various domains.

- Email: sarani@cmu.ac.ir
- ORCID: 0000-0001-6598-2410
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://www.cmu.ac.ir/staff/sarani

Rasoul Damani received the B.Sc, M.Sc and Ph.D. degrees from Sharif University of Technology (SUT), Tehran, Iran in 1998, 2000 and 2015 respectively, all in Electrical Engineering. He is currently an Assistant Professor at Chabahar maritime university, Chabahar, Iran. His research interests include the areas of optical communication and underwater communication systems.

- Email: damani@cmu.ac.ir
- ORCID: 0000-0002-4748-0684
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://www.cmu.ac.ir/staff/rdamani

How to cite this paper:

Z. Raisi, V. M. Nazarzehi Had, E. Sarani, R. Damani, “FATR: A comprehensive dataset and evaluation framework for persian text recognition in wild images,” J. Electr. Comput. Eng. Innovations, 13(2): 331-340, 2025.

DOI: 10.22061/jecei.2024.11256.784

URL: https://jecei.sru.ac.ir/article_2253.html