IDPL-PFOD: An Image Dataset of Printed Farsi Text for OCR Research

Fatemeh sadat Hosseini
Intelligent Data Processing Laboratory,
Department of Electrical Engineering, Shahid Bahonar University of Kerman,
Kerman, Iran
ftmsdt98@gmail.com

Elham Shabaninia
Department of Computer Engineering, Sirjan University of Technology,
Sirjan, Iran
eshabaninia@sirjantech.ac.ir

Shima Kashef
Faculty of Sciences and Modern Technologies, Graduate University of Advanced Technology,
Kerman, Iran
sh.kashef@kgut.ac.ir

Hossein Nezamabadi-pour
Intelligent Data Processing Laboratory,
Department of Electrical Engineering, Shahid Bahonar University of Kerman,
Kerman, Iran
nezam@uk.ac.ir

Abstract

The existence of appropriate image datasets in the field of optical character recognition (OCR) plays an essential role in the accuracy of OCR systems. Despite the fact that many image datasets with different richness have been published to date, the Farsi (Persian) image datasets are very few. Also, there is a shortage of image datasets that contain sentences or lines of real text. Although Farsi and Arabic have many similarities, the differences between the two scripts cause the OCR systems trained with Arabic datasets, do not have proper accuracy on Farsi texts. The main purpose of the present article is to introduce a Printed Farsi Dataset for OCR researches (call as IDPL-PFOD). This dataset is made from Miras text dataset and the images are generated with different fonts, font styles, font sizes and backgrounds. Also, to increase the similarity of the generated images to the real images some blur and distortion have been added to the images.

1 Introduction

In recent centuries, people have devoted the early years of their lives to learning how to read and write, and after learning these abilities, they have found an acceptable ability to read handwritten and printed texts in a variety of fonts. It can be said that some people have the ability to read texts printed in fancy fonts or texts written in fonts known as calligraphic fonts. Despite all these advances and researches that have been done for nearly 5 decades, the reading ability of computers is still far behind the ability of humans (Naz et al. 2014). Therefore, empowering computers to read different texts was considered by researchers. Optical character recognition is the task of recognizing the existing texts from images and scanned documents and converting them to a text file that has the ability to search and edit on the computer (Kashef 2021; Singh, Bacchuwar, and Bhasin 2012; Nanehkaran et al. 2021). OCR is used in a wide range of fields. These applications include diagnosis of biomedical science (Nanehkaran et al. 2021), handwritten Farsi digits (Nanehkaran et al. 2021), banking (Ganis, Wilson, and Blue 1998), health care industry (Ganis, Wilson, and Blue 1998), captcha (Gossweiler, Kamvar, and Baluja 2009), institutional repositories and digital libraries (Barwick 2007), optical music recognition (Singh et al. 2011), automatic number plate recognition (Pandey et al. 2017; Kashef, Nezamabadi-pour, and Rashedi 2018; Rakhshani 2019), handwritten recognition (Plamondon and Srihari 2000; Arani, Kabir, and Ebrahimpour 2019), reading and verifying bank checks (Naz et al. 2014), verifying people's signature (Hafemann, Sabourin, and Oliveira 2017), etc.

The first step to do all the research in the field of OCR in any language is to collect a dataset with a sufficient number of samples and appropriate variety to be able to provide a realistic environment for the OCR system (Mozaffari et al. 2008). In addition, the existence of a standard dataset plays an essential role in the development, testing and comparison of different recognition systems and helps researchers to evaluate and compare their recognition techniques. Therefore, it can be concluded that a standard dataset can play an important role in promoting researches (Mozaffari

et al. 2008; Safabaksh, Ghanbarian, and Ghiasi 2013; Memon et al. 2020). Research on OCR has been conducted in many languages including English, Arabic, Hindi, Chinese, Korean, Urdu, Farsi (Memon et al. 2020; Kashef 2021). Despite extensive advances of OCR in the English language, other languages, especially Farsi, have lagged. Ziaratban, Faez, and Bagheri (2009) consider one of the reasons for the backwardness of the Farsi language in comparison with the English language to be the inherent features of the Farsi language.

Datasets used in OCR research have several classifications based on where they are used and how they are generated.

- Generally, there are three types of OCR datasets including handwritten, printed and scene-text (Torabzadeh and Safabaksh 2015). Handwritten and printed datasets are respectively created by photographing or scanning handwritten and printed texts, and scene-text datasets are created by photographing or scanning photos containing text with a complex or patterned background.
- From the perspective of how the dataset is generated, OCR datasets are divided into real and artificial ones. Real datasets are created by scanning documents and images that contain text, but artificial datasets are created from ready-made texts and have the ability to use a variety of fonts, noise, and backgrounds. They are usually images of each line or word in a text (Kashef 2021).

Based on the above explanations and considering that official documents need to be recognized these days, in this article we intend to introduce a Printed Farsi Dataset for Farsi optical character recognition researches, IDPL-PFOD (IDPL stands for "Intelligent Data Processing Laboratory" and PFOD stands for "Printed Farsi OCR Dataset"). As far as we know, there is no artificial dataset of printed texts in Farsi whose images contain text lines with proper variety in font, size and style. Our dataset contains 30,138 images, each image contains a line from Miras text dataset[1] which is a Farsi news corpus. To generate these images, we used common Farsi fonts and font styles, several font sizes and different backgrounds such as plain white, textures and noisy to increase diversity. Also, a portion of images is created with some distortion and blur that are usually seen in scanned texts.

This paper is organized as follows. Features of the Farsi script and related works are discussed in Section 2 and the steps for creating IDPL-PFOD are explained in Section 3. In section 4, IDPL-PFOD characteristics are discussed and the paper is concluded in Section 5.

2 Background

2.1 Farsi Language

One of the branches of Indo-Iranian languages is Persian or Farsi, which is the official language of Iran, Afghanistan and Tajikistan. Also, some people in Uzbekistan speak Farsi. Farsi is the second most widely spoken language in Southwest Asia and has been introduced as the language of culture and education in several Muslim countries. Although Memon et al. (2020) mentions that Farsi script is similar to Arabic, Urdu, Pashto and Dari languages, it has significant differences with other Indo-Iranian languages, especially Arabic (Haghighi et al. 2009). Therefore, the Farsi language needs its own datasets, and it should be noted that the best results for a recognition system are obtained when a suitable dataset is used. According to Torabzadeh and Safabaksh (2015), none of the previous datasets are comprehensive enough to satisfy all the parameters needed for Farsi text recognition systems.

Before the main features of a suitable dataset for Farsi texts are introduced, features of the Farsi language are discussed in the following:

a. Farsi language has 32 characters which are written from right to left (Haghighi et al. 2009).

b. Several pairs of characters in Farsi have a similar appearance, and the only difference in these characters is the number and position of dots (Azmi and Kabir 1999). For instance:

(ب، پ، ت) ، (چ، ج، ح، خ) ، (د، ذ)
، (ر، ز، ژ) ، (س، ش) (ص، ض) ، (ط، ظ)
(ع، غ) ، (ک، گ)

c. Farsi is a cursive script. In other words, its characters attach to each other in writing (Safabaksh, Ghanbarian, and Ghiasi 2013; Haghighi et al. 2009; Solimanpour, Sadri, and Suen 2006).

d. Each Farsi character may have a maximum of 4 writing styles depending on its position in the

[1] https://github.com/miras-tech/MirasText

word; beginning style, middle style, ending style and isolated style (Safabaksh, Ghanbarian, and Ghiasi 2013; Torabzadeh and Safabaksh 2015; Haghighi et al. 2009; Azmi and Kabir 1999), For instance:

beginning style: صـ ، بـ ، نـ ، فـ ، خـ ، لـ ، مـ ، عـ

middle style: ـصـ ، ـبـ ، ـنـ ، ـفـ ، ـخـ ، ـلـ ، ـمـ ، ـعـ

ending style: ـص ، ـب ، ـن ، ـف ، ـخ ، ـل ، ـم ، ـع

isolated style: ص ، ب ، ن ، ف ، خ ، ل ، م ، ع

See this example for more clarity for the letter "ع". It may form combinations like these: beginning style: "علی", middle style: "بعد", ending style: "بدیع", isolated style: "شجاع".

e. As implied in the 4th case, each Farsi character, depending on its position in the word, can be connected to other characters from one or both sides, and some of them can be written separately, and this causes "sub-words" to appear. In fact, sub-words are the parts of a word that are written separately (Torabzadeh and Safabaksh 2015; Kashef 2021; Azmi and Kabir 1999; Solimanpour, Sadri, and Suen 2006), For instance: word "صابون" can be separated into three sub-words "صا" ,"بو" ,"ن".

f. Some Farsi characters can be written in different styles when we use different fonts (Jaiem et al. 2013; Solimanpour, Sadri, and Suen 2006), For instance: Sein character with:

a) IranNastaliq font written like: "س"

b) Nazanin font written like: "س"

g. The other thing to know about the Farsi alphabet is that it does not have capital letters. This means that both proper names and ordinary nouns are written with the same letter forms.

h. Although Farsi and Arabic are very similar (Safabaksh, Ghanbarian, and Ghiasi 2013), but they have some differences.

a) Arabic has 28 characters, but Farsi has 32 characters. In other words, Farsi has 4 characters more than Arabic including: "پ (p)", "ژ (zh)", "گ (g)" and "چ (ch)" (Safabaksh, Ghanbarian, and Ghiasi 2013; Haghighi et al. 2009).

b) Digits 4, 5 and 6 have different styles in Farsi and Arabic scripts. The style of writing the digits 4, 5 and 6 in Farsi is "4", "5" and "6", but in Arabic is "4", "5" and "6" respectively.

c) Some characters in Arabic have no use in Farsi, like: "ة" and "ي". Also, some punctuation marks are less commonly used in Farsi writing, like: "ـــ" ,"ـِـ" ,"ـُـ".

i. Depending on the level of literacy and geographical area, Persian speakers write some numbers in several ways (Nanehkaran et al. 2021).

j. In Farsi, numbers are written from left to right like in English, although in Farsi, words, sentences and dates are written from right to left (Solimanpour, Sadri, and Suen 2006; Azmi and Kabir 1999).

According to the aforementioned, it is necessary to have a dataset that is specific for the Farsi language and has the proper features of a dataset used for OCR researches. Ref. (Torabzadeh and Safabaksh 2015) lists the main features of the appropriate dataset used for Farsi OCR researches as follows:

- Having real words: The image on which the recognition operation is performed are usually scanned documents that contain noise, distortion and blur depending on the quality of the scan. So, the appropriate dataset should contain real words with these features.
- Having a uniform distribution of characters in the language under study: In the training phase of optical character recognition systems, if we want to train each character fairly, the number of instances of each character must be almost equal. However, as we all know, every language has both repetitive characters and characters that are less commonly used. In Farsi, the two characters "ا" and "ی" are very frequent and the two characters "ء" and "ظ" are infrequent.
- Having all the characters of the language under study: As mentioned, Farsi has 4 characters more than Arabic, so we can't only rely on Arabic datasets to train good Farsi recognition systems.
- Supporting common fonts of the researched language: Most real Farsi documents are written in common Farsi fonts. In order to increase the accuracy of recognition systems, it is necessary to use common Farsi fonts. Whereas the existing Arabic datasets have used fonts that are very different from the widely used Farsi fonts.
- Including different font sizes: Recognition systems have made their performance depends on the font size. Many of these systems can only be compatible with large fonts. Therefore, in order for the recognition system to perform better, although the recognition process becomes more complex, it is better to use different font sizes.

2.2 Related works

Although the existence of Latin printed datasets has reached an acceptable maturity, the lack of datasets in other languages, especially Farsi, is felt. In the following, published datasets for Farsi, Arabic and Urdu languages that have been published since 2009 are reviewed. In fact, to the best of our knowledge, only 2 Farsi image datasets (Torabzadeh and Safabaksh 2015; Asadi 2020) for printed text, have been published to date, but almost an acceptable number of Arabic datasets of printed texts have been published. However, because the Farsi, Arabic and Urdu scripts are similar in most features, we also review them.

APTI: APTI stands for "Arabic Printed Text Image", which was published in 2009 (Haghighi et al. 2009). APTI is an artificial dataset whose words are taken from a dictionary of 113,284 words and has 45,313,600 images. Each image contains a printed Arabic word that is generated from 10 fonts, 10 sizes and 4 styles. The final point about this dataset is the existence of ground truth data in XML file format provided for the dataset and this dataset is available to the public[2] .

PATDB: PATDB stands for "Printed Arabic Text Database" (Al-Hashim and Mahmoud 2010). This dataset is published in 2010 and contains scanned images of various Arabic printed texts such as chapters of books, advertisements, magazines, newspapers and reports with resolutions of 200 or 300 or 600 dpi. Total images in this dataset are 6954 pages scanned, and it is publicly available.

APTID/MF: APTID/MF stands for "Arabic Printed Text Image Database/Multi-Font", is published in 2013 (Jaiem et al. 2013). In this dataset, 387 pages of printed Arabic documents have been scanned and finally, 1,845 blocks of text have been obtained with grayscale format and 300 dpi resolution. Also, it contains 27,402 Arabic printed character images. The dataset and its ground truth data provided for it are available to the public.

UPTI: UPTI stands for "Urdu Printed Text Image Database", is published in 2013 (Sabbour and Shafait 2013). This dataset is similar to the APTI dataset and has more than 10,000 images of Urdu text which are written in Nastaliq font.

AUT-PFT: This dataset is published in 2015 and has 10,000 words using 127 unique characters. Words in this dataset are meaningless because the distribution of all characters is the same throughout the dataset. In this dataset, all generated images are printed and scanned to add real noise to the images. It is worth mentioning that the words written in the pictures with 10 widely used Farsi fonts and 4 different font sizes. The bottom line about this dataset is the existence of ground truth data in XML file format provided for the dataset (Torabzadeh and Safabaksh 2015).

ALTID: ALTID stands for "Arabic/Latin Text Image Database", is published in 2015 (Chtourou et al. 2015). In this dataset 731 pages of printed and handwritten Arabic and Latin documents have been scanned and finally, 1,845 Arabic text and 2,328 Latin text images have been obtained. The images are in grayscale and with a resolution of 300 dpi. The ground truth data is also provided for this dataset.

Smart ATID: ATID stands for "Arabic Text Image Dataset", is published in 2016 (Chabchoub et al. 2016). This dataset contains images for Arabic handwritten and printed documents captured by mobile phones in different conditions (blur, perspective angle and light). This dataset has two groups of images including printed and handwritten documents. The first group, which is close to our aim, is prepared from 116 paper documents and a total of 16,472 document pages have been created. The ground truth data is also provided for the dataset.

PATD: PATD stands for "Printed Arabic Text Dataset", and is published in 2019 (Bouressace and Csirik 2019). In this dataset, 810 images are scanned by mobile phone in different conditions (blur, different perspective angle and different light), with the grayscale format and different resolutions. It has been extracted from 10 newspapers which are written in 14 fonts and 3 font styles. finally, 2,954 images are created with multi-fonts, multi-font sizes and multi-font styles.

Shotor: The latest version of this dataset is published in 2020 and has 120,000 grayscale 50*100 images, each of which has a meaningful Farsi word written in different fonts and font sizes. In fact, there are 120,000 meaningful Farsi words extracted from Farsi Wikipedia[3] and Ganjoor[4]

[2] https://diuf.unifr.ch/main/diva/APTI/
[3] https://fa.wikipedia.org
[4] https://ganjoor.net

Name	Language	Samples	Type of samples	Plain white	Noisy	Texture
ALTID	Arabic/Latin	1,845/2,328	Document pages	✓	✓	×
PATDB	Arabic	6954	Document pages	✓	✓	×
APTID/MF	Arabic	27,402	Document pages	✓	✓	×
Smart ATID	Arabic	16,472	Document pages	✓	✓	×
PATD	Arabic	810	Document pages	✓	✓	×
APTI	Arabic	45,313,600	Words	✓	✓	×
UPTI	Urdu	10,000	Ligatures	✓	✓	×
AUT-PFT	Farsi	10,000	Meaningless words	✓	✓	×
Shotor	Farsi	120,000	Words	✓	×	×
IDPL-PFOD	Farsi	30,138	Lines (part of sentences, 452,070words)	✓	✓	✓

Table 1: Summary of Arabic, Urdu and Farsi datasets

site. This dataset is publicly available (Asadi 2020)[5].

Table 1 shows a summary of Related works.

3 Building the IDPL-PFOD

In this section, the steps for creating the IDPL-PFOD dataset are explained.

3.1 Text Processing

In order to have the first feature of a suitable dataset which is having real words, Miras text dataset (Sabeti et al. 2018) is selected. Miras and Hamshahri (Baradaran Hashemi, Shakery, and Faili 2010) text datasets, contain news text and are very popular and available to the public. As far as we know, the size of the Miras dataset is much larger than Hamshahri, therefore several articles recognizing Farsi texts use Hamshahri dataset. To increase diversity and innovation, in this dataset, we have used Miras text dataset with a volume of more than 200 GBs.

Miras text dataset contains 2,835,414 news items, and 753 news of this text dataset is used in IDPL-PFOD. Since some of this news is non-Farsi, in the first step of text processing, non-Farsi news is removed. Doing this, 643 news remains, out of 753 selected news. As Miras text dataset contains other information in addition to the original news text, such as news sources, and the publisher's website, the unnecessary information is removed in the second step of text processing to only have the original news text. In the third step, some bad characters, such as © and ﷺ have been removed, because these characters are not defined for Farsi fonts. Due to the use of the previous steps in the processed text, a significant number of unnecessary "space" characters are created. Therefore, to have a clean text, it is necessary to remove these unnecessary space characters. This is done in the fourth step of text processing. In the fifth step, the character "ي" is replaced with the character "ی" because the character "ي" is not used in Farsi, but was mistakenly used instead of the character "ی" in the text. Finally, because some of the lines are very long, in the last step a standard line length is defined. We considered this standard length to be 15 words. Thus, 30,138 lines were created as final lines. In total, this dataset has 452,070 words.

3.2 Image Generation

To write any line generated in the previous step on an image, we must first specify the fonts, font sizes, and font styles. Therefore, in the following, first, the above items about IDPL-PFOD are mentioned and then the image generation process is discussed. For the Farsi language, like other languages, many fonts with different styles are designed. However, not all of these fonts are widely used and most of them are used for graphical works. In this paper, we select 11 fonts out of 39 fonts that have been modified and standardized by the Iran Supreme Council of Information and Communication Technology (SCICT) under Standard No. 6219 which is published on its website[6] . According to SCICT, out of 11 selected fonts, 10 of them are the most commonly used fonts in Farsi texts and printed documents. In addition, due to the high use of the "Titr" font in official documents, we have also used the "Titr" font to generate images. The 11 selected fonts are Badr, Compset, Lotus, Mitra, Nazanin, Roya, Traffic, Titr, Yagut, Yekan and Zar.

[5] https://github.com/amirabbasasadi/Shotor

[6] www.scict.ir

Recent research has shown that more than one style per font is used to increase the variety of datasets. In this study, we also used two styles for most fonts. There are two styles including Bold and Normal for all fonts except for the two fonts “Titr” and “Yekan” that only have the bold and normal styles, respectively. Also, due to the uniform use of all fonts and uniform distribution, 30,138 lines should be divided into 11 equal parts, but since 30,138 is not divisible by 11, so 2,740 images are generated for the first 9 fonts (alphabetically) and 2,739 images are generated for the remaining 2 fonts. In order to use all the styles of each font, the font style is selected randomly for each image from the available styles. Recent research has also shown that it is useful to use multi-font sizes to add variety to the dataset to be closer to real data. According to our research, the font size of real documents is between 10 and 16, so we have randomly selected the font size of the texts written on the images as a random number in this range. Based on the above and what will be explained below, we have started to generate images by using the Python programming language and have saved them with numbers 1 to 30,138 in “tif” format. The generated images of each font are saved in a folder called the name of that font. It should be noted that the images are similar in terms of dimensions which are 700*50 pixels.

3.3 Add Background to image

This image dataset is generated with the aim of being used in training different OCR systems. For this purpose, we use three types of backgrounds to enable researchers to use this dataset as printed and scene text datasets. The three types of backgrounds we have used are plain white, noisy and texture. We have used 4 types of noise to generate images with a noisy background including Gaussian, Speckle, Salt and Pepper and Poisson noises. Also, 12 different patterns are used to generate textured images. It is already mentioned that almost the same number of images is generated for each font. The three backgrounds are used for images of each font (folder) with the frequency of 50% for plain white, 40% for noisy, and 10% for texture backgrounds. In other words, for each font 1,370 images with plain white background, 1,096 images with noisy background, and 273 (274 for the first 9 fonts) images with textured background have been generated. Therefore, a total of 2,740 or 2,739 images were generated for each font. By performing a simple calculation, it is concluded that in this dataset, there are 15,070 images with plain white back background, 12,056 images with noisy background, and 3,012 images with textured background.

Figures 1-3 show three examples of generated images with different backgrounds, fonts and font sizes.

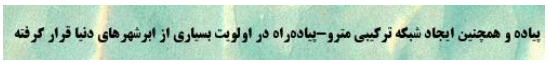

Figure 1: Background: Texture, Font: B Titr Bold, Font size: 14, Distortion: None, Blur: None.

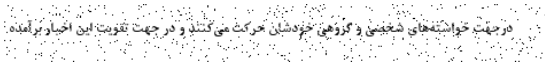

Figure 2: Background: Noisy (S&P), Font: B Nazanin, Font size: 13, Distortion: None, Blur: None.

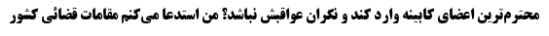

Figure 3: Background: Plain white, Font: B Titr Bold, Font size: 16, Distortion: None, Blur: None.

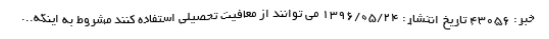

Figure 4: Background: Plain white, Font: B Yekan, Font size: 13, Distortion: Sinewave, Blur: None.

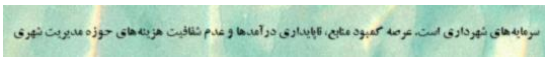

Figure 5: Background: Texture, Font: B Zar Bold, Font size: 13, Distortion: None, Blur: Gaussian.

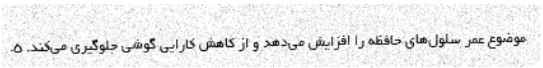

Figure 6: Background: Noisy(gaussian), Font: B Yekan, Font size: 16, Distortion: Sloping (1 degree), Blur: None.

Figure 7: Background: Texture, Font: B Yagut, Font size: 13, Distortion: Sloping (-1 degree), Blur: Gaussian.

3.4 Add Blur and Distortion to image

Since the quality of real images is not always excellent, to bring the generated images closer to the real images, a bit of sloping distortion (-1 degree or 1 degree) or sinewave distortion (to add sinewave distortion, we have used part of the code published on GitHub[7]) and gaussian blur is added randomly to the generated images. The exact statistics are as follows: 4% sloping distortion, 1% sinewave distortion, 3% blur, 2% both blur and one type of distortion.

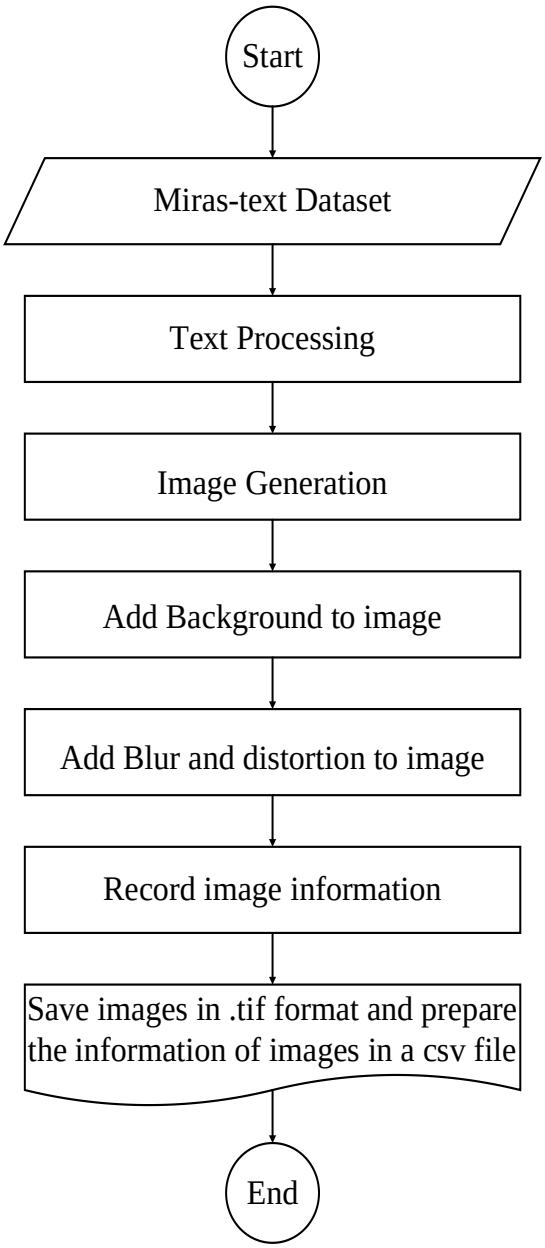

Figure 8: The flowchart of steps for creating the IDPL-PFOD dataset

[7]https://github.com/Belval/TextRecognitionDataGenerator

Figures 4-7 show images with different backgrounds, fonts, font sizes, distortions and blur.

3.5 Record image information

According to the description, each image can have one of the 11 fonts with different styles and sizes. Also, the background type can be one of the 3 mentioned models, and even the image may contain distortion or blur. Therefore, to access information of each image, a CSV file is created and all this information is saved in it. This CSV file contains 7 columns and 30,138 rows, each row belonging to an image. The first to fourth columns of this file shows the following information, respectively: the name of the image, background type, font with the font style and font size. The fifth column reports whether the image is distorted or not, and if it is distorted, sloping distorted (to which degree) or sinewave distorted is specified. The sixth column records whether the image is blurry or not and finally, the last column reports the correct text used in the image.
Figure 8 shows the flowchart of steps for creating the IDPL-PFOD.

4 Discussion

The main features of a suitable dataset that is used for Farsi recognition systems are mentioned in Section 2. Here are some of these features about IDPL-PFOD.

Due to the use of Miras text dataset, which is created from real news, the text data used in IDPL-PFOD has real words. Therefore, IDPL-PFOD has the first feature. In real texts, some characters are more common in any language, so when we use real texts, we can no longer maintain a uniform distribution of characters. Therefore, IDPL-PFOD does not have the second feature. Although the distribution of characters in real texts is not equal, but all characters are presented in real texts. So, our dataset also has the third feature. As mentioned, we have used the fonts introduced by the SCICT as the most widely used fonts, so the fourth feature is also available in IDPL-PFOD. Note that the range of fonts used in the IDPL-PFOD is the range of fonts in administrative documents. Therefore, the last feature is also available in IDPL-PFOD. In addition to the above features, IDPL-PFOD also simulates the actual environment conditions of a printed text such as distortion, blur, noisy and

texture backgrounds. In this paper, we list all the steps of data generation in detail, including text processing, fonts, font size, font style, type of noise, blur, and distortions. All image information along with the ground truth text is saved in a CSV file. IDPL-PFOD is open to the public with the following two links:

GitHub link:
https://github.com/FtmsdtHosseini/IDPL-PFOD

IDPL website link:
https://idpl.uk.ac.ir/%D8%AF%DB%8C%D8%AA%D8%A7-%D8%B3%D8%AA

Fonts	Font styles	Font sizes
11	2	7

Table 2: No. fonts, font styles, font sizes

Texture	Distortion	Noise	Blur
12	3	4	1

Table 3: No. texture, distortion, noise, blur

	Plain white	Noisy	Texture	Total image
Each font	1,370	1,096	273/ 274	2,739/ 2,740
All font	15,070	12,056	3012	30,138
Total Lines	30,138			
Total words	452,070			

Table 4: No. images in each background and each font

Tables 2-4 show a summary of IDPL-PFOD dataset information.

5 Conclusion

The accuracy of OCR systems highly depends on the dataset selected for the training phase. Since the creation of real datasets requires many problems, including cost and manpower, artificial datasets can be an appropriate alternative to real datasets. Therefore, generating datasets with appropriate richness and near-realistic samples is very necessary for this branch of research. To date, many image datasets with different richness have been published for different languages. However, despite that Farsi is the second language of the Southwest Asian continent, there are very few Farsi image datasets. Although the number of Arabic datasets is large, and Farsi OCR systems can be trained with Arabic datasets due to the similarities of the two scripts, the differences of these scripts reduce the accuracy of Farsi OCR systems. Therefore, the need for a rich Farsi database is strongly felt. Also, there is no Farsi image dataset that includes sentences or lines of real text. In this paper, we introduce a Printed Farsi Dataset (IDPL-PFOD) for Farsi optical character recognition researches, which is prepared from the 753 news out of 2,835,414 news items of the Miras text dataset. We generate 30,138 images out of 753 selected news with eleven fonts: Badr, Compset, Lotus, Mitra, Nazanin, Roya, Traffic, Titr, Yagut, Yekan and Zar, with two randomly font styles Normal and Bold, and seven font sizes which are randomly selected between 10 and 16. Also, three backgrounds are used including plain white, noisy (Gaussian, Speckle, Salt and Pepper and Poisson noises) and texture (12 different patterns randomly used). Also, to increase the similarity of the generated images with real images, a little blur (Gaussian blur) and distortion (sloping distortion or sinewave distortion) have been added to the images randomly. To access the information of each image, a CSV file is created which includes the name of the image, background type, used font and the font style, font size, whether the image is distorted or not and whether the image is blurry or not. It should be noted that all steps are done in Python, which is the most widely used programming language for Data Science research, and are available to the IDPL website and IDPL-PFOD repository.

References

Al-Hashim, Amin G, and Sabri A Mahmoud. 2010. "Printed Arabic text database (PATDB) for research and benchmarking." In *Proceedings of the 9th WSEAS international conference on Applications of Computer Engineering*, 62-68. Citeseer.

Arani, Seyed Ali Asghar Abbaszadeh, Ehsanollah Kabir, and Reza Ebrahimpour. 2019. 'Handwritten Farsi word recognition using NN-based fusion of HMM classifiers with different types of features', *International Journal of Image and Graphics*, 19: 1950001.

Asadi, Amir Abbas. 2020. "Shotor Dataset." In. https://www.kaggle.com/amir137825/persianocrdataset/version/2.

Azmi, R., and E. Kabir. 1999. 'A New Segmentation Technique for Omnifont Farsi Text', *iut-jame*, 18: 1-10.

Baradaran Hashemi, Homa, Azadeh Shakery, and Heshaam Faili. 2010. "Creating a Persian-English Comparable Corpus." In *Multilingual and Multimodal Information Access Evaluation*, edited

by Maristella Agosti, Nicola Ferro, Carol Peters, Maarten de Rijke and Alan Smeaton, 27-39. Berlin, Heidelberg: Springer Berlin Heidelberg.

Barwick, Joanna. 2007. 'Building an institutional repository at Loughborough University: some experiences', *Program*, 41: 113-23.

Bouressace, Hassina, and Janos Csirik. 2019. "Printed Arabic Text Database for Automatic Recognition Systems." In *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, 107–11. Istanbul, Turkey: Association for Computing Machinery.

Chabchoub, F., Y. Kessentini, S. Kanoun, V. Eglin, and F. Lebourgeois. 2016. "SmartATID: A Mobile Captured Arabic Text Images Dataset for Multi-purpose Recognition Tasks." In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 120-25.

Chtourou, I., A. C. Rouhou, F. K. Jaiem, and S. Kanoun. 2015. "ALTID : Arabic/Latin Text Images Database for recognition research." In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 836-40.

Ganis, M. D., C. L. Wilson, and J. L. Blue. 1998. 'Neural network-based systems for handprint OCR applications', *IEEE Transactions on Image Processing*, 7: 1097-112.

Gossweiler, Rich, Maryam Kamvar, and Shumeet Baluja. 2009. "What's up CAPTCHA? a CAPTCHA based on image orientation." In *Proceedings of the 18th international conference on World wide web*, 841–50. Madrid, Spain: Association for Computing Machinery.

Hafemann, Luiz G., Robert Sabourin, and Luiz S. Oliveira. 2017. 'Learning features for offline handwritten signature verification using deep convolutional neural networks', *Pattern Recognition*, 70: 163-76.

Haghighi, Puntis Jifroodian, Nicola Nobile, Chun Lei He, and Ching Y. Suen. 2009. "A New Large-Scale Multi-purpose Handwritten Farsi Database." In *Image Analysis and Recognition*, edited by Mohamed Kamel and Aurélio Campilho, 278-86. Berlin, Heidelberg: Springer Berlin Heidelberg.

Jaiem, Faten Kallel, Slim Kanoun, Maher Khemakhem, Haikal El Abed, and Jihain Kardoun. 2013. "Database for Arabic Printed Text Recognition Research." In *Image Analysis and Processing – ICIAP 2013*, edited by Alfredo Petrosino, 251-59. Berlin, Heidelberg: Springer Berlin Heidelberg.

Kashef, S., H.Nezamabadi-pour, and E.Shabani. 2021. 'An Overview of Deep Learning Methods in Optical Character Recognition with Emphasis on Persian, Arabic and Urdu Calligraphy', *Machine vision and image processing*.

Kashef, Shima, Hossein Nezamabadi-pour, and Esmat Rashedi. 2018. 'Adaptive enhancement and binarization techniques for degraded plate images', *Multimedia Tools and Applications*, 77: 16579-95.

Memon, J., M. Sami, R. A. Khan, and M. Uddin. 2020. 'Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)', *IEEE Access*, 8: 142642-68.

Mozaffari, Saeed, Haikal El Abed, Volker Märgner, Karim Faez, and Ali Amirshahi. 2008. "IfN/Farsi-Database: a database of Farsi handwritten city names." In *International Conference on Frontiers in Handwriting Recognition*.

Nanehkaran, Y. A., Defu Zhang, S. Salimi, Junde Chen, Yuan Tian, and Najla Al-Nabhan. 2021. 'Analysis and comparison of machine learning classifiers and deep neural networks techniques for recognition of Farsi handwritten digits', *The Journal of Supercomputing*, 77: 3193-222.

Naz, S., A. I. Umar, S. B. Ahmed, S. H. Shirazi, M. Imran Razzak, and I. Siddiqi. 2014. "An Ocr system for printed Nasta'liq script: A segmentation based approach." In *17th IEEE International Multi Topic Conference 2014*, 255-59.

Pandey, Atul, Vivek Sharma, Shruti Paanchbhai, Neha Hedaoo, and SD Zade. 2017. 'Optical character recognition (OCR)', *International Journal of Engineering and Management Research (IJEMR)*, 7: 159-61.

Plamondon, R., and S. N. Srihari. 2000. 'Online and off-line handwriting recognition: a comprehensive survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22: 63-84.

Rakhshani, S., E.Rashedi,H.Nezamabadi-pour. 2019. 'Number plate recognition using deep learning', *Machine vision and image processing*.

Sabbour, Nazly, and Faisal Shafait. 2013. "A segmentation-free approach to Arabic and Urdu OCR." In *Document recognition and retrieval XX*, 86580N. International Society for Optics and Photonics.

Sabeti, Behnam, Hossein Abedi Firouzjaee, Ali Janalizadeh Choobbasti, SHE Mortazavi Najafabadi, and Amir Vaheb. 2018. "Mirastext: An automatically generated text corpus for persian." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Safabaksh, R., A. R. Ghanbarian, and G. Ghiasi. 2013. "HaFT: A handwritten Farsi text database." In *2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP)*, 89-94.

Singh, A., K. Bacchuwar, A. Choubey, D. Kumar, and S. Karanam. 2011. "An OMR based automatic music player." In *2011 3rd International Conference on Computer Research and Development*, 174-78.

Singh, Amarjot, Ketan Bacchuwar, and Akshay Bhasin. 2012. 'A survey of OCR applications', *International Journal of Machine Learning and Computing*, 2: 314.

Solimanpour, Farshid, Javad Sadri, and Ching Y. Suen. 2006. "Standard Databases for Recognition of Handwritten Digits, Numerical Strings, Legal Amounts, Letters and Dates in Farsi Language." In *Tenth International Workshop on Frontiers in Handwriting Recognition*. La Baule (France): Suvisoft.

Torabzadeh, S., and R. Safabaksh. 2015. "AUT-PFT: A real world printed Farsi text image dataset." In *2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP)*, 267-72.

Ziaratban, M., K. Faez, and F. Bagheri. 2009. "FHT: An Unconstraint Farsi Handwritten Text Database." In *2009 10th International Conference on Document Analysis and Recognition*, 281-85.