



A novel comprehensive database for offline Persian handwriting recognition



Javad Sadri ^{a,b,*}, Mohammad Reza Yeganehzad ^b, Javad Saghi ^b

^a Department of Computer Science & Software Engineering, Faculty of Engineering and Computer Science, Concordia University, Montreal, Quebec, Canada, H3G 1M8

^b Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, P.O. Box: 615/97175, Birjand, Iran

ARTICLE INFO

Article history:

Received 15 April 2015

Received in revised form

31 January 2016

Accepted 16 March 2016

Available online 29 March 2016

Keywords:

Persian handwriting recognition

Persian offline recognition

Persian handwriting database

Check recognition

Numerical string

Digit recognition

Persian date

Persian alphabet

Unconstrained Persian handwriting recognition

ABSTRACT

Developing a standard database for offline handwriting recognition is an essential task. This paper offers a novel comprehensive database for conducting research on offline Persian handwriting recognition. Seven pages of forms were designed and completed by 500 native Persian writers, who were equally balanced in terms of gender and randomly selected from all over Iran. Then, the completed forms were scanned at a resolution of 300 DPI. Through several intensive processing steps, a huge number of isolated digits, numeral strings, touching digits, dates, words, names, alphabetical letters, free texts, arithmetic, and especial symbols from all these forms were extracted and organized as a standard database. All samples in this database were assigned with detailed ground truth and stored in three color formats: true color, gray level, and binary. Also, all subsets of this database were randomly partitioned into training, validation, and testing sets. We hope this comprehensive database will extend research in the pattern recognition community.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Offline handwriting recognition refers to the ability of a machine to receive and interpret a previous individual-made handwritten input from a photographed or scanned image [1]. Offline handwriting recognition has many real world critical applications: Recognition of postal addresses used in mail sorting [2], calculation of handwritten arithmetic expressions [3], and the processing of handwritten bank checks [4] are only a few well-known examples of these applications. Another significant application of offline handwriting processing is word spotting used to search for a keyword or phrase amongst a large number of handwritten documents [5]. Researchers working on handwriting recognition applications believe that without utilizing a large standard handwriting database, none of the performance factors of a system can be evaluated or improved [6]. Throughout common scripts, research on Persian script recognition suffers from the same problem, which is the lack of a large comprehensive handwriting

database for Persian [7]. Although Persian script has many similarities and tight connections to Arabic script, their letters, words, and styles of writing are not exactly the same [8]. Hence, none of the Arabic handwriting databases can be effectively used for the purpose of Persian script recognition. This paper presents a novel comprehensive benchmark of Persian handwriting database aiming to alleviate difficulties in offline Persian handwriting recognition and to expand research in all aspects of Persian script recognition.

Persian as an ancient language has evolved over centuries and is currently the official language of more than 110 million people, most of whom live in countries such as Iran, Tajikistan, Afghanistan, and other neighboring countries [8]. Persian script has some features that distinguish it from other scripts. For example, the Persian alphabet letters are an extension and slightly modified version of Arabic letters, and they have similarities to Dari, Urdu, and Pashto letters. Handwritten words in Persian and other Arabic-related scripts are not the same, however they do share the same writing direction (right to left), and their words are written in cursive. Automatic recognition of Persian handwritten script is very difficult; some challenges in offline Persian handwriting recognition are as follows:

* Corresponding author at: Department of Computer Science & Software Engineering, Faculty of Engineering and Computer Science, Concordia University, Montreal, Quebec, Canada, H3G 1M8. Tel.: +1 514 848 2424x3000; fax: 514 848 2830.

E-mail addresses: j_sadri@encs.concordia.ca (J. Sadri), m.yeganehzad@gmail.com (M.R. Yeganehzad), saghii.ac@gmail.com (J. Saghi).

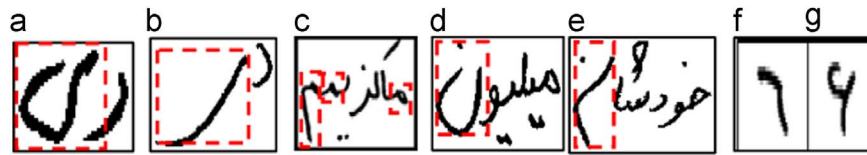


Fig. 1. Some challenges of Persian handwriting recognition: (a and b) cursiveness of the character "ی" in the word "دی" ("Dei"), (c) different shapes of the letter "م" in the word "ماکزیم" ("maximum"), (d), formal writing style of the character "ن" written at the end of the word "میلیون" ("million"), (e) the same character in the same position within another word "خودشان" ("themselves") written by another writing style, and (f and g) two different writing styles of digit "۶", '6'.

- Generating cursiveness while writing words: for example, the letter "ی" (pronounced "yeh") in the word "دی" (a Persian month name, pronounced 'Dei'). See Fig. 1a and b.
- Different shapes of letters with respect to their positions in words: for example, the letter "م" (pronounced "mim") has three distinct shapes in the word "ماکزیم" ("maximum"). See Fig. 1c.
- Writing some letters in different shapes apart from their formal shapes: for example, as seen in Fig. 1d and e, the letter "ن" (pronounced "noon") in the word "خودشان" ("themselves") is written differently from its formal shape in the word "میلیون" ("million").
- Different writing styles of some digits, such as "۶", '6', see Fig. 1f and g.

Taking these challenges into consideration, providing a comprehensive handwriting database is a crucial priority for research and future developments of Persian handwriting recognition systems.

The lack of a comprehensive database for Persian handwriting recognition has led to some researchers creating their own handwriting databases for their publications [9,10]. Most of the efforts for creating standard databases were on Latin [11–13], Chinese [14], Korean [15], Japanese [16], Indian [17] and Arabic [18,19] scripts. There have been a few databases for Persian, such as CENPARMI¹ [7], including isolated digits, dates, isolated letters, and legal amounts. Additionally, Ziaratban et al. offered an unconstrained Persian handwritten text database [20]. Likewise, Khosravi and Kabir provided a database of binary level Persian digit images [21]. While these works were great endeavors toward the creation of offline Persian handwriting database, several important deficiencies were perceived, such as a lack of touching digits, samples of punctuations/symbols, popular words, names, legal amounts, worded dates, sentences, free handwritten texts, a lack of equal distribution of men and women writers, and a lack of providing various image modalities. Hence, we were inspired to alleviate these shortcomings. The main contributions of this paper are two-fold: firstly, to create a generalized unified and large-scale yet comprehensive handwritten collection set in three different color modalities covering: dates, digits, numeral strings of variable length, touching digits, letters, symbols, words, and an unconstrained text that includes all of the aforementioned items. The samples were collected from 500 writers (250 males and 250 females) chosen randomly from Iran's various geographical regions, with varying literacy levels, 11–53 years of age. Secondly, a generalized unified framework for creating comprehensive handwriting databases was introduced, which can be followed by researchers working on handwriting recognition of any scripts. The database introduced in this paper provides very good opportunities for conducting researches on topics such as skew correction, line segmentation/detection, numeral string detection/segmentation, word spotting/segmentation/recognition, writer identification, gender detection, and handedness detection, as well as

for conducting experiments on machine learning, pattern recognition, image processing, feature extraction approaches, and developing practical Persian handwriting recognition systems. In order to show the usefulness of our database in research, several experiments were conducted in this paper.

The rest of the paper is organized as follows: in Section 2 the data collection process is explained, in Section 3 the process of data extraction and preparation is described, in Section 4 the database structure and statistics is overviewed, in Section 5 some possible applications and conducted experiments are described. A comparison of our database with similar works in other languages is discussed in Section 6 and conclusions and future works are discussed in Section 7.

2. Data collection

As shown in Fig. 1, Persian script has many unique characteristics which present some challenges in handwriting recognition [8]. In order to cover all these characteristics and challenges a set of 500 native Persian writers, equally balanced in terms of gender (250 males, 250 females) were randomly selected from all over Iran. Among them, left-handed writers constituted almost 10% (22 males and 31 females). The set also covered different literacy levels, from primary school to post-secondary educations, and an extensive age range, from 11 to 53 years. Writers filled out specially designed forms with various fields and layouts to capture a complete picture of Persian script along with sufficient samples of its different written items such as: dates, digits, numeral strings, letters, words, names, and texts. As shown in Fig. 2, our standard seven-page data-entry forms were designed intuitively, including entry fields with typewritten labels and unlabeled ones. The entry fields were of an adequate size such that the majority of writing styles fit within their boundaries. Also, on the corners of each form, four small black boxes were designed in order to simplify de-skewing of the forms as well as locating the information on them after the scanning process. During completion of the forms no limitation on writing style or writing instrument was imposed. Selection of numerous distinct writers and design of the forms' pages created proper conditions to capture all possible challenges in Persian handwritten script. In the next sections, contents and layout of each page of our forms are briefly described.

2.1. Structure of Page 1

Layout of Page 1 was divided into 2 blocks: header block and data entry block. A filled out instance of Page 1 is shown in Fig. 2a. In the header block, the writer's ground truth information, viz. name, family name, gender, handedness, age, and education level were collected. In the data entry block, isolated digits and numeral strings with lengths of 2, 3, 4, and 5 were included.

2.2. Structure of Page 2

Numeral strings of 6, 7, or 10 digits' length were included in Page 2 (see Fig. 2b). In order to enhance legibility of numeral

¹ Center for Pattern Recognition and Machine Intelligence.

Figure 2 displays the layout of 7 pages (a-g) of data collection forms. The forms are organized into sections: Header Block and Data-Entry Block. Page (a) shows the header and data entry sections. Page (b) shows the header and data entry sections with fields F1, F2, and F3 highlighted. Page (c) shows the header and data entry sections with field F1 highlighted. Page (d) shows the header and data entry sections with field F1 highlighted. Page (e) shows the header and data entry sections. Page (f) shows the header and data entry sections with field F1 highlighted. Page (g) shows the header and data entry sections with field F1 highlighted.

Fig. 2. Layout of 7 pages (a–g) of our data collection forms that were completed by all 500 writers.

strings, in two 7-digit strings two commas were used (see field F1 in Fig. 2b). Due to the abundant use of 10-digit strings in national and postal codes, they were incorporated in Page 2. F2 fields in Fig. 2b show Persian dates in numeral and worded forms. Likewise, the participants wrote the names of boys, girls, and cities freely as shown in F3 fields.

2.3. Structure of Page 3

As shown in Fig. 2c, 32 Persian alphabetical letters in 4 different forms, viz. isolated, initial, medial, and final, covered more than 2/3 of the fields of Page 3. In the F1 fields of Fig. 2c, popular Arabic letters used in Persian scripts are shown. Page 3 was ended with common arithmetic symbols.

2.4. Structure of Page 4

Page 4 was started with some of the widely used punctuation symbols in Persian handwritings (see F1 fields in Fig. 2d). The rest of the Page 4 entries were devoted to common Persian words.

2.5. Structures of Pages 5 and 6

Worded numbers used in financial sheets covered the entirety of Page 5, as seen in Fig. 2e, and part of the Page 6 entries, as shown by F1 in Fig. 2f. Page 6 was ended with Persian and Arabic month names.

2.6. Structure of Page 7

In Page 7 of the form, participants wrote a fixed typewritten text to form one of the novel contributions of the current research. Standard text properties were put together in the provided text: dates, numeral strings, cursive words, and different city names beside special symbols. Having masked the writer's name due to

confidentiality issues, a handwritten text sample of Page 7 is shown in Fig. 2g. The four black corner boxes for de-skewing the forms are highlighted in this figure.

In total, 3500 ($=500 \times 7$) Handwriting Sample Forms (HSFs) were collected, and scanned in true color (24 bits per pixel) with 300 DPI resolution and were stored in TIF standard format files. Each handwriting sample file of each writer was assigned a unique name, for example: ID0000107_P1.tif or ID0000107_P2.tif (for Pages of 1 and 2 of the form written by writer ID#=107, respectively). In the next section, the data extraction process used on these HSFs is explained.

3. Data extraction

After scanning the HSFs, their handwriting samples were extracted. The data extraction process was conducted in two main steps. The details of these steps are shown in a flowchart in Fig. 3 and are explained briefly in the two following sections.

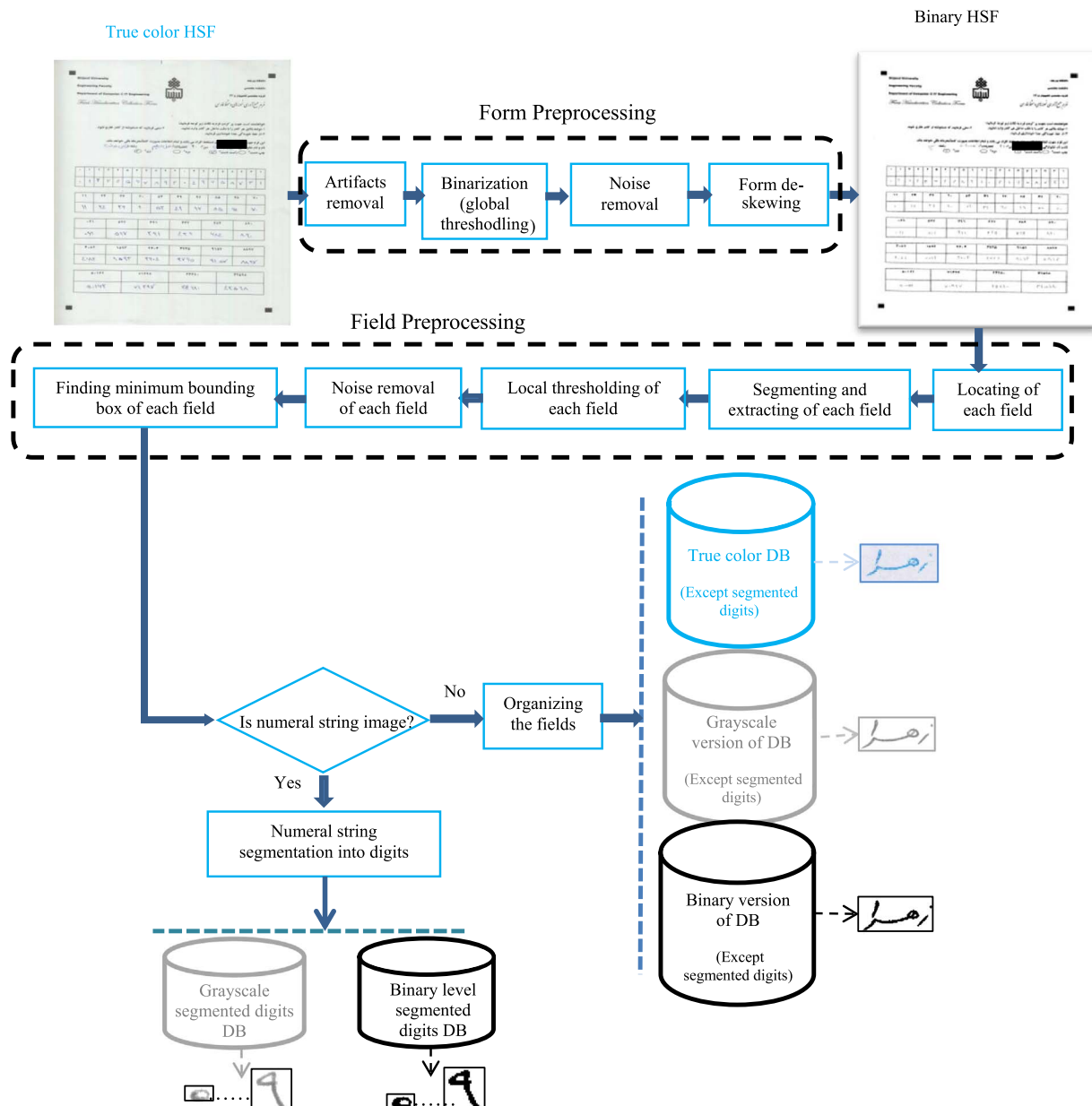


Fig. 3. Overall view of the cleaning and data extraction process from our HSFs, and creation of 3 versions of our database including: true color, gray level, and binary.

3.1. Form preprocessing

As seen in the flowchart of Fig. 3, the preprocessing step consisted of four stages carried out in order to create proper conditions for cutting out images of handwritten fields from the scanned HSF. Removing probable artifacts generated by scanning (such as the black border in Fig. 2f) was the first stage. In the second stage, a binary version of the form was derived from the true color version by applying the Otsu method [22]. Then, median filtering [23] was applied to remove salt and pepper noise on the binary form. De-skewing of the HSF was the fourth preprocessing stage. Estimation of skew angle (θ) and skew correction was simply handled by horizontal and vertical aligning of the four black corner boxes designed on each HSF (see Fig. 2g).

3.2. Field preprocessing

Pursuant to Fig. 3, a database was constructed after preprocessing the forms. After skew correction of the true color form using θ , coordinates of the handwritten fields were found. These coordinates were applied to the corresponding true color HSF to segment and extract each true color field on the form. In the next step, for each true color extracted field, a gray threshold was calculated using the gray-level histogram [22]. This local calculated threshold is fine-tuned manually, so as to improve the quality of the resulting gray field (through an eye-verification process). On the resulting gray image, components with length and width of less than 5 pixels were removed as noise. In the last step, a minimum bounding box of the resulting gray image for each field was employed on the corresponding true color field for trimming. Ultimately, on the gray field, all pixels with brightness level less than the adjusted threshold were set to black and the rest to white to create a binary version of each field. As seen in Fig. 3, true color, grayscale, and binary versions of the database were created and organized accordingly.

Further processing was required to create a database of segmented digits from gray numeral strings. This step was conducted by segmenting digits of the bounded numeral strings through connected component analysis (see Fig. 4 for more details).

4. Database overview

As mentioned in previous sections, a comprehensive handwriting database including samples of all written items in Persian has been provided. The overall view of this comprehensive database beside all its different folders and all subsets of its images is shown in Fig. 5. As seen in this figure, our comprehensive database is comprised of eight main datasets (or folders) of written items in Persian handwritten script including: Dates (numeric and

worded), Digits, Numeral Strings, Touching Digits, Alphabetical Letters, Punctuation Marks and Symbols, Words, and unconstrained free Texts. Additionally, all HSFs (500×7), 500 samples per page, used to collect these items are made available in the database. As shown in Fig. 5, all the images of this database have been prepared in three formats: true color, gray level, and binary in order to provide the researchers with more options for their experimentations. All the images in all folders have been divided randomly into three separate and distinct datasets: training (60% of all images), validation (20% of all images), and testing (20% of all images). This database also contains the ground truth data and meta-data which contain the detailed description of all image contents and their writers' information (such as: ID, age, handedness, gender) in all folders. More details about our ground truth information will be explained in Section 4.9. More elaborate explanations of all items and different folders of this database are provided in the following sections.

4.1. Dates database

The writing format of Persian dates, the so-called "Hijri-Shamsi" calendar system, is **year/month/day**. Each participant wrote five numeral and four worded dates freely to cover an extensive range of numeral and worded dates. Fig. 6 shows two binary samples of Persian dates. In total, 2500 numeral dates and 2000 worded dates were collected.

4.2. Digits database

In Persian, similar to Latin, digits and numeral strings are written from left to right. In our database each digit (0–9) is written 20 times by each participant ($20 \times 10 = 200$ digits per writer overall). In order to investigate various shapes and writing styles of digits, each participant wrote digits in different locations in the data collection forms: twice in isolated format and eighteen times within numeral strings. It was noticed in our collected samples that when a digit is written in a numeral string, its size and slant is affected by its location. Furthermore, it might touch other digits, overlap with neighboring ones, or even have a shape that differs from the isolated format. In Fig. 7, examples of handwritten numeral strings collected from the same participant are shown. The participant wrote all the three numerals started by '۶' with the shape of "۶" in Fig. 7a, whereas the same digit in the middle position of the leftmost numeral has been written as "6". Also in Fig. 7b, the collected isolated style of digit '6', written in the form of "۶", is shown (both styles of writing '6' are acceptable in Persian). Before adding the digits to the database all of them were verified manually and if a scratched digit was found, it was added to another folder called 'garbage collection.' For example, in Fig. 7c there is a case where

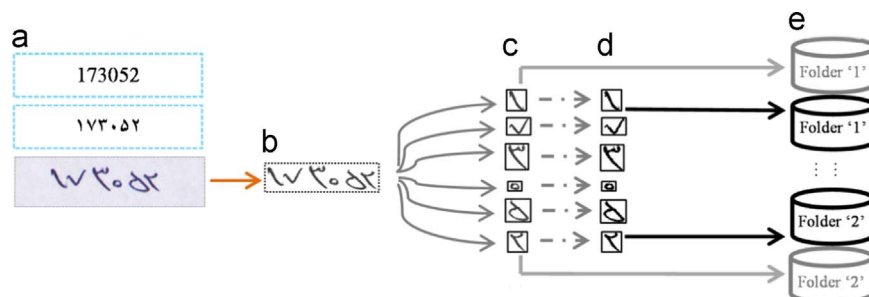


Fig. 4. Segmented digits database construction procedure; (a) a cut out numeral string image with their typewritten equivalent in Persian and English, (b) the bounded grayscale image, (c) grayscale version of the segmented digit, (d) binary version of the segmented digits. (e) The obtained digit images were stored in the corresponding gray-level and binary folders of our database.

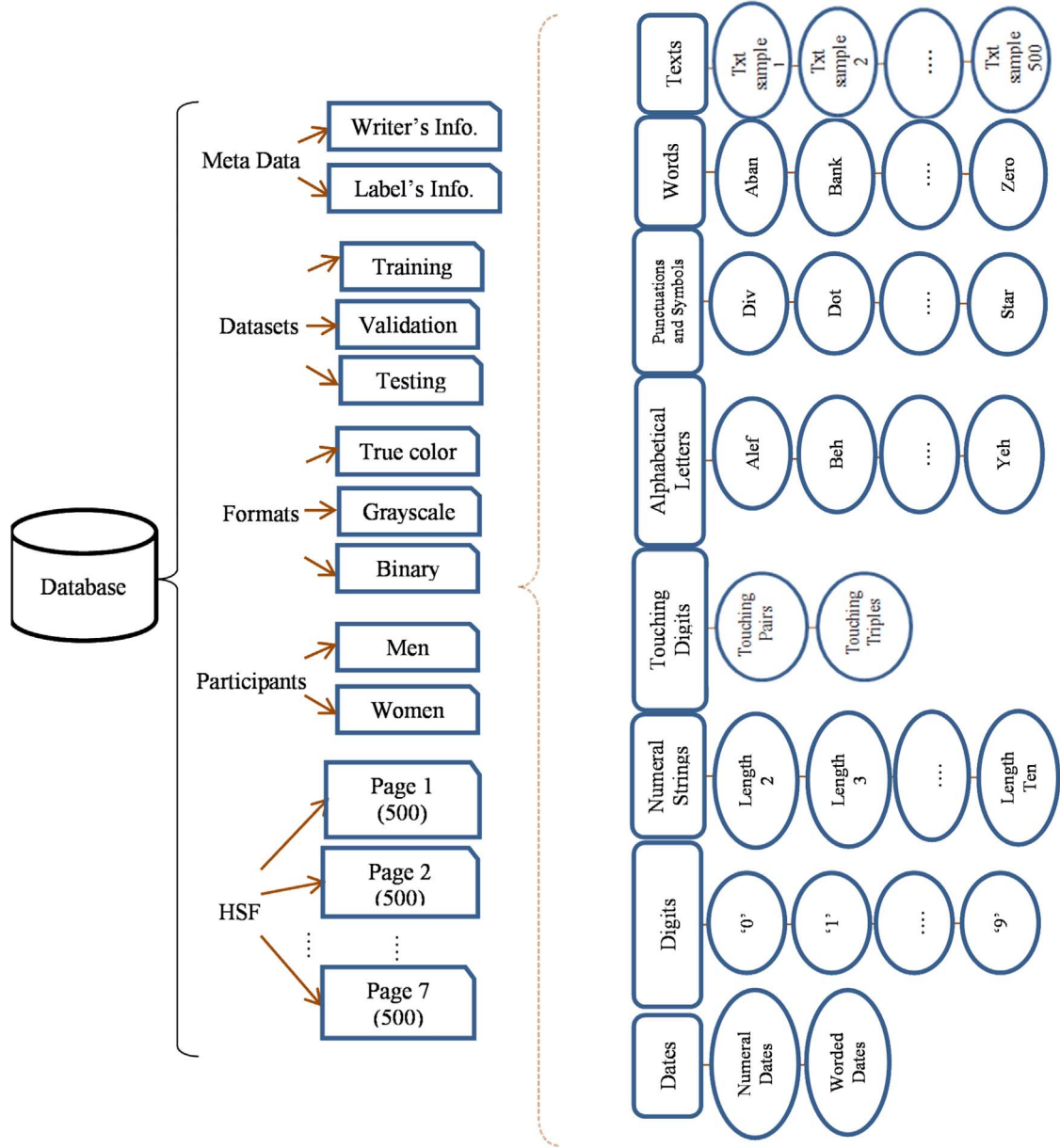


Fig. 5. Database structure overview.



Fig. 6. Two examples of Persian handwritten dates based on Hijri-Shamsi calendar with their typewritten equivalent in Persian and English: (a) a numeral date (b) a worded date.

the participant scratched while writing the digit '7'. This scratched digit was added to 'garbage collection.' In addition, some participants mistakenly wrote digits in numeral strings, such as the digit '4' in Fig. 7d. These cases had carefully been specified in the eye verification step before assigning their corresponding ground truth. Statistics of our digit database after these corrections are tabulated in Table 1. As seen, this database has been balanced for training and testing sub-folders.

4.3. Numeral strings database

In our data collection forms, each participant wrote 39 numeral strings of varying length. The number of each numeral string of predetermined length gathered from each participant is illustrated in Fig. 8.

Each digit (0–9) had been distributed intuitively within various-length numeral strings in the data collection forms shown

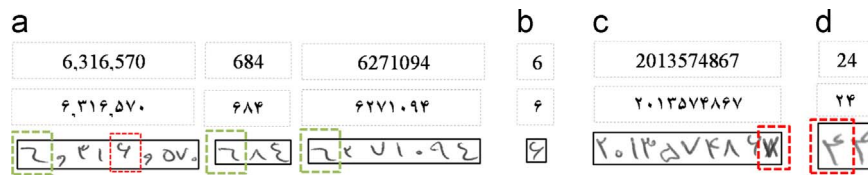


Fig. 7. Persian handwritten digit samples in terms of isolated form and numeral string: (a) various writing shapes of digit '6' in two distinct positions within numeral strings, (b) digit '6' in the isolated form (c) the scratched digit '7' (d) the participant's mistake in writing '4' instead of '2'.

Table 1
Persian digits database statistics.

Digit	Total	Training set	Validation set	Testing set
0	9700	6000	1700	2000
1	9758	6000	1758	2000
2	9770	6000	1770	2000
3	9733	6000	1733	2000
4	9623	6000	1623	2000
5	9756	6000	1756	2000
6	9733	6000	1733	2000
7	9718	6000	1718	2000
8	9640	6000	1640	2000
9	9693	6000	1693	2000
0–9	97,124	60,000	17,124	20,000

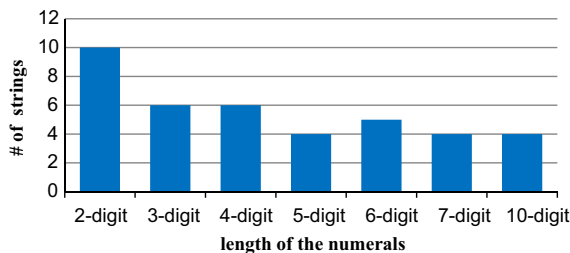


Fig. 8. Number of numeral strings with predefined lengths collected from each participant.

in Fig. 2. In Fig. 9, each digit placement percentage within numeral strings is shown. For example, in Fig. 9a it is reported that out of 39 numeral strings, each participant started two numeral strings with the digit '0' ($2/39 \approx 5\%$). In Fig. 9b, it is shown that the digit '0' was the last written digit in 5 cases ($5/39 \approx 13\%$). Accordingly, across 102 digits written in the middle of the strings (excluding 2×39 starting and ending digits), each writer wrote '0' in the middle 11 times ($11/102 \approx 11\%$) as noted in Fig. 9c. Hence, the problem of digit variation because of writing speed was alleviated by collecting all Persian digit images in various positions within numeral strings. In Fig. 10a, a pair of 2-digit and 7-digit strings collected from the same writer is compared. The effect of the speed of writing has been revealed by elongating the tail of the left string digits. In Fig. 10b two '0's are compared: the left one is similar to a circle (with an empty interior) whereas the right one, as the ending component of the string, is closer to a dot. Also, elongation of the tail of the digit '6' in the left string compared to the right one is discerned. These cases were collected from one writer, while many other interesting writing styles of digits within numeral strings were found and saved in the database. As a result we have a very rich collection of shapes and styles of written numerals. Statistics of our numeral strings' database are shown in Table 2.

4.4. Touching digits database

In spite of accomplished researches on segmentation and recognition of touching digits for Latin numeral strings [24,25], so far there was no dataset for touching cases in Persian. As a result,

segmentation of Persian touching digits is an open problem. Therefore, in addition to isolated digits and numeral strings, we have also created a touching digits folder in our database. All touching digits found on our collected forms were stored in this folder. In total, 450 touching pairs and five touching triples were extracted from the numeral strings dataset. In Fig. 11a and b two instances of touching pairs and touching triples are shown respectively.

4.5. Alphabet letters database

Persian script includes 32 alphabetical letters which is an extension of Arabic letters plus four Persian-specific letters [8]. In Persian and Arabic scripts, in addition to various writing styles of the shapes of the letters, their formal glyphs differ with respect to their positioning in a word. Therefore, the initial, medial, final, and isolated forms of Persian letters were collected from each writer. In Table 3, some handwritten letters in all possible forms written by a participant have been captured.

There was a punctual investigation of the preprocessed images of the handwritten letters' database. Looking into the collected data, there were forms of letters written in other distinct glyphs rather than the ones in Table 3, for example: the letter "ه" (pronounced "heh") was also written "هـ" and "هـ" in its medial and final forms, respectively. The letter "س" (pronounced "sin") was also written "سـ". Moreover, the letter "ا" (without diacritics) (pronounced "alif") can represent several phonemes: 'hamza' above, "أ", 'hamza' under, "إ", with 'maddah', "آ", with 'tanvin' above, "اَ" (pronounced "an"). Despite having been adopted from Arabic texts, they are very common in Persian script. Likewise, "ک" and "ق" (pronounced "keh"), are the same letters with different formal writing forms, however "ق" is more popular in Arabic texts. Similarly, "ی" and "ی" have the same description so they are summarized in one row.

Also, occasionally, two or more special letters are joined as a single glyph when they are adjacent. This typography is called ligature [26]. Two very popular ligatures in Persian (and Arabic) texts are: 'lam+alif', "لا" (pronounced "laa") and 'lam+alif+tanvin', "لاَ" (pronounced "lan"). Different writing styles of these ligatures are available in the current database but are sometimes ignored in Persian or Arabic handwriting databases. Furthermore, there are some Arabic glyphs, collected from writers: "لا", called 'ta marbutah,' "ى", the combination of 'ya+hamza' beside its initial form, and "و", the combination of 'waw+hamza.' While these glyphs are originally Arabic they are widely used in Persian scripts, therefore they have been included in our database. In all, 86 glyphs (letter shapes) per writer were collected in the alphabet letter database. The overall statistics of the isolated letters database are shown in Table 4.

4.6. Punctuation and symbols database

We have also captured Persian symbols and punctuations in our database. This section consists of two types of symbols: punctuation symbols and basic arithmetic expression symbols, explained in following sections.

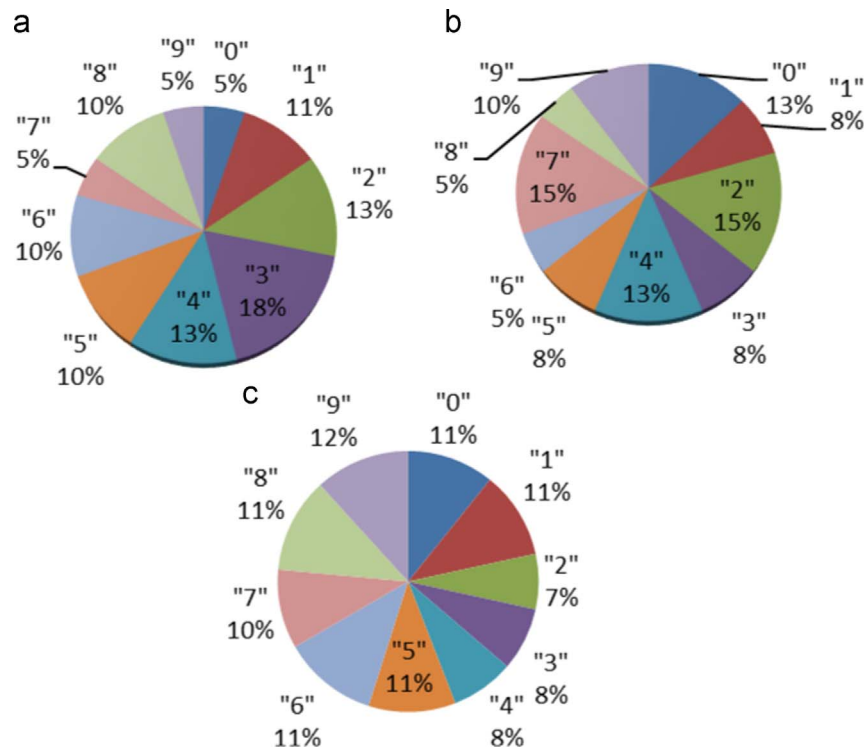


Fig. 9. Each Persian digit placement percentage within the numeral strings (a) at the beginning, (b) at the end, and (c) in the middle of the string.

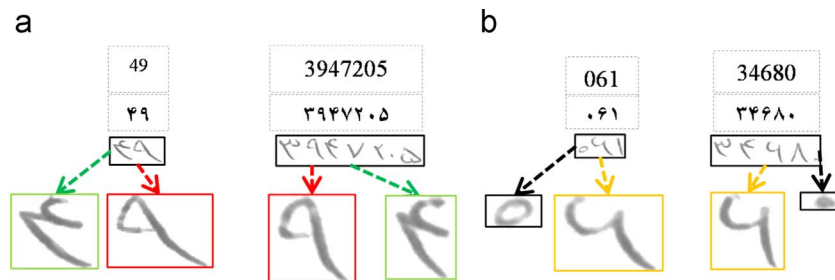


Fig. 10. Comparison of numeral strings' digits in various lengths written by the same individual: (a) string length of 2 and 7, and (b) string length of 3 and 6.

4.6.1. Punctuation symbols

From among tens of punctuation symbols, 23 of the most well-known were chosen and collected from each participant. As shown in Table 5, three Persian-specific punctuations including the comma, semi-colon, and question mark, plus the Latin comma (","), were collected. The rest of the collected punctuations were not specific to any language and were gathered to examine the Persians' writing style.

4.6.2. Basic arithmetic symbols

Basic handwritten arithmetic expressions share another part of the punctuation and symbols database. An example set of 9 handwritten arithmetic expressions collected from a writer is tabulated in Table 6. This part of our database can be used for research on arithmetic expression recognition surrounding Persian handwritings.

It should be noted that the majority of the special symbols database can also be used for Arabic, Dari, Urdu, Pashto, and other scripts. Statistics of the available images in this database are shown in Table 7.

Table 2

Numeral strings database statistics.

Numeral strings' length	Total	Training set	Validation set	Testing set
Two	5000	3000	1000	1000
Three	3000	1800	600	600
Four	3000	1800	600	600
Five	2000	1200	400	400
Six	2500	1500	500	500
Seven	2000	1200	400	400
Ten	2000	1200	400	400
Two-Ten	19,500	11,700	3900	3900

4.7. Words database

Handwritten words in Persian are written in cursive forms and normally they have various writing styles. In our database, we have tried to gather a rich collection of common Persian words and their writing styles. Some samples of the collected words are shown in Fig. 12. Based on the type of usage in conventional contexts, these words are categorized as follows: (1) legal and natural person titles, including 5 words, (2) commercial product items, including 10 words, (3) ordinal and cardinal numbers

including 36 words, (4) legal amount words which are used in bank checks, including 50 words (5) names of months of the year in Persian; since there are two calendar systems used in Persian (solar and lunar calendars, known as “Hijri-Shamsi” and “Hijri-Ghamari,” respectively). The total number of month names in our database is 24 ($=12+12$). (6) Handwritten names, including boys’ and girls’ first names (10 samples) plus 5 Persian city names. In total, each of our 500 participants wrote 140 different words, distributed amongst these categories as explained above. Statistics of this word database are shown in Table 8.

4.8. Texts database

Persian free handwritten text is another unique contribution of our present Persian database. A paragraph of handwritten text was designed and collected in order to contain and show variations of digits, numeral strings, letters, symbols, words, and sentences within the context of a paragraph, and enrich their corresponding datasets. Each participant wrote a sample text freely including 271 words, 6 numeral strings, and some punctuation symbols embedded in the text which showed the variation of these items in the context of a paragraph, lines, and other surrounding words. This text has provided a good opportunity for researching problems such as: text line segmentation/extraction, baseline

detection/correction, segmentation and extraction of numeral strings on text lines from surrounding words, word segmentation, word spotting, word recognition, slant correction, line skew correction, etc. in the area of handwriting recognition. To the best of our knowledge, no similar database has included items such as numeral strings and words in an unconstrained, free format text. In Fig. 13, a sample of handwritten text with its corresponding ground truth content is shown. The writer’s name has been masked for confidentiality. This text database has also been partitioned into a training set (300 samples), a validation set (100 samples), and a testing set (100 samples).

4.9. Metadata and ground truth

Metadata and ground truth files are one of the most essential parts of our database. Such a huge database with different collections and thousands of items and samples should contain very

Table 4
Statistics of the isolated letters database.

Total	Training set	Validation set	Testing set
43,000	25,800	8600	8600

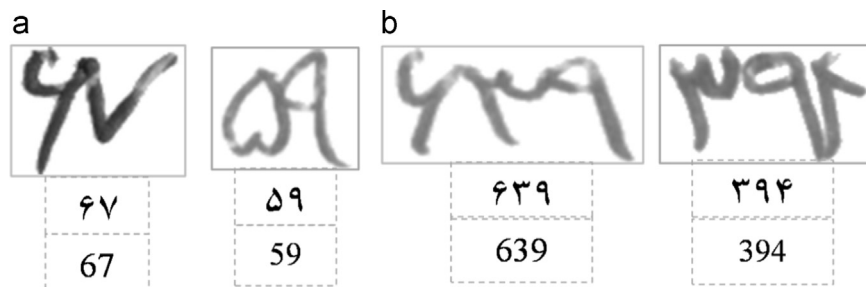












Fig. 11. Touching digits samples: (a) touching pairs and (b) touching triples.

Table 3
Some handwritten alphabet letters in terms of isolated, initial, medial and final forms beside their general names.

General name		Isolated form		Initial form		Medial form		Final form	
Tcheh ^{PS}		چ		چ		چ		چ	
Seen		س		س		IF	IF	IF	IF
Ain		ع		ع		ع		ع	
Kehe		ک		ک		IF	IF	IF	IF
Heh		ه		ه		ه		ه	
Yeh	Arabic Ya	ی		ی		ی		IF	IF

PS: Persian-specific, not used in Arabic scripts.
IF: similar to isolated form and not collected.

Table 5
A sample of punctuation symbols collected in our database with their names.










Unicode name	Symbol	Handwritten sample	Unicode name	Handwritten sample	Written sample
Comma ^{PS}	,		Single left-pointing angle quotation mark	◀	◀
Semicolon ^{PS}	;		Double right-pointing angle quotation mark	»	»
Comma	,		Double left-pointing angle quotation mark	«	«
Exclamation mark	!		Right bracket]]
Question mark ^{PS}	?		Left bracket	[[
Hyphen-minus	-		Right parenthesis))
Underline	—		Left parenthesis	((
Ellipsis	...		Right Brace	}	}
Dot	.		Left Brace	{	{
Colon	:		Single right-pointing angle quotation mark	›	›

PS: Persian-specific.

rich metadata to facilitate using and interpreting image samples. To this end, each image sample in this database had to be characterized with full image identity and writer information. Accordingly, metadata in the present work were composed of two Ground Truth (GT) files: image ground truth and writer ground truth, each stored in text (.TXT) and Microsoft Excel[®] file formats. The structures of these files are shown in Table 9a and b, respectively. As shown in Table 9a, the GT attributes of each image sample include image specifications such as: image file name, writer ID, field type (dates, digits, numeral strings, touching digits, isolated letters, special symbols, words or texts), sub-field name, content, partition type (training (R), validation (V), or testing (T)), and image type (true color, gray level, or black/white). As shown in Table 9b, the writer's GT includes the writer's specifications, such as: writer ID, gender, handedness, age, educational level, and field of study. Also, an XML file has been provided as an additional GT data for texts database in order to simplify working with text samples in terms of finding words and sub-words in different lines and paragraphs.

As seen in Table 9a, each image sample of the database has a unique image file name. These image file names were derived and enriched through an organized process such that they could be traced back to their original places in their corresponding HSFs (see Fig. 14). If we categorize the database samples into non-digit items and digit items, the naming procedure for corresponding files was similar with a trivial difference. There were four sections in the non-digit image filename strings (five sections for digit items) separated by the “_” character and each section contained specific numbers: the first section corresponded to the participant ID, the second section to the page number, the third to the box number (the boxes were numbered from left to right and from top to bottom as in Fig. 14a), and the fourth to image type (true color (TC), binary (BW), or gray level (GL)) (see Fig. 14b, c, and d). For digit items, a new section was required to address the location of digits within the original numeral strings where these digits had been segmented. This new section shows the connected component (CC) number in the segmented numeral string (see Fig. 14e).

Table 6
Basic arithmetic symbols written by a participant.

General name	Symbol	Written sample
Backslash	\	
Obelus	÷	
Percent	%	
Plus	+	
Multiplication sign	×	
Equality	=	
Slash	/	
Divisor		
Asterisk	*	

5. Applications and some experimental results

Researchers working in the domains of image processing and handwriting recognition can set up their experiments based on the current database. Accordingly, some of the prominent industrial applications that can be developed using our database are listed in the next section. Afterwards, some experiments are conducted to show the applicability of our database.

5.1. Applications

As can be seen in Fig. 5, our database has several handwritten collections of items frequently used in Persian script, all of which were written by 500 Persian writers (250 men, 250 women). These collections include a huge number of image samples (in true color, gray level, and binary formats) and their associated ground truths, including: dates (worded and numeric), isolated digits, numeral strings, touching digits, isolated alphabet letters, punctuation marks and symbols, words, and free texts. Samples in these collections can be used for solving research problems in handwriting

recognition, specifically in Persian script recognition, and for conducting various studies such as: handwritten date recognition, digit recognition, numeral string segmentation, mathematical expressions, letter recognition, text line segmentation/extraction, baseline detection/correction, segmentation and extraction of numeral strings on text lines from surrounding words, word segmentation, word spotting, word recognition, slant correction, normalization, binarization of digits and words, line skew correction, writer identification, gender detection, handedness detection, testing supervised, semi-supervised, unsupervised learning or clustering methods, image processing, feature extraction, segmentation methods, and important applications such as check recognition, mail sorting, creating general Persian handwriting recognition softwares, etc.

5.2. Experimental results

Among all important research fields and commercial applications discussed above, this section is dedicated to an example to show how our database can be used for isolated digit recognition

Table 7
Statistics of the special symbols database.

Total	Training set	Validation set	Testing set
16,000	9600	3200	3200

Table 8
Words database statistics.

	Total	Training set	Validation set	Testing set
Words	70,000 (= 140 × 500)	42,000	14,000	14,000

a			b		
خانم	شرکت	موسسه	ماشین	اتومبیل	تلفن
خانم	شرکت	موسسه	ماشین	اتومبیل	تلفن
خانم	شرکت	موسسه	ماشین	اتومبیل	تلفن
Lady	Corporation	Institution	Machine	Automobile	Telephone
c			d		
پنجم	نهمصد	چهارده	ریال	معادل	حامل
پنجم	نهمصد	چهارده	ریال	معادل	حامل
پنجم	نهمصد	چهارده	ریال	معادل	حامل
Fifth	Nine hundred	Fourteen	Rial	Equal to	Bearer
e			f		
ربیع الثانی	ذی القعدة	آبان	پیمان	هانیه	ارومیه
ربیع الثانی	ذی القعدة	آبان	پیمان	هانیه	ارومیه
ربیع الثانی	ذی القعدة	آبان	پیمان	هانیه	ارومیه
Rabie-o-sani	Zel-ghade	Aban	Amirhossein	Samira	Mashhad

Fig. 12. Persian words database samples: (a) legal and natural persons, (b) commercial product items, (c) cardinal and ordinal numbers, (d) Hijri-Shamsi and Hijri-Ghamari month names, and (e) individual and city names.

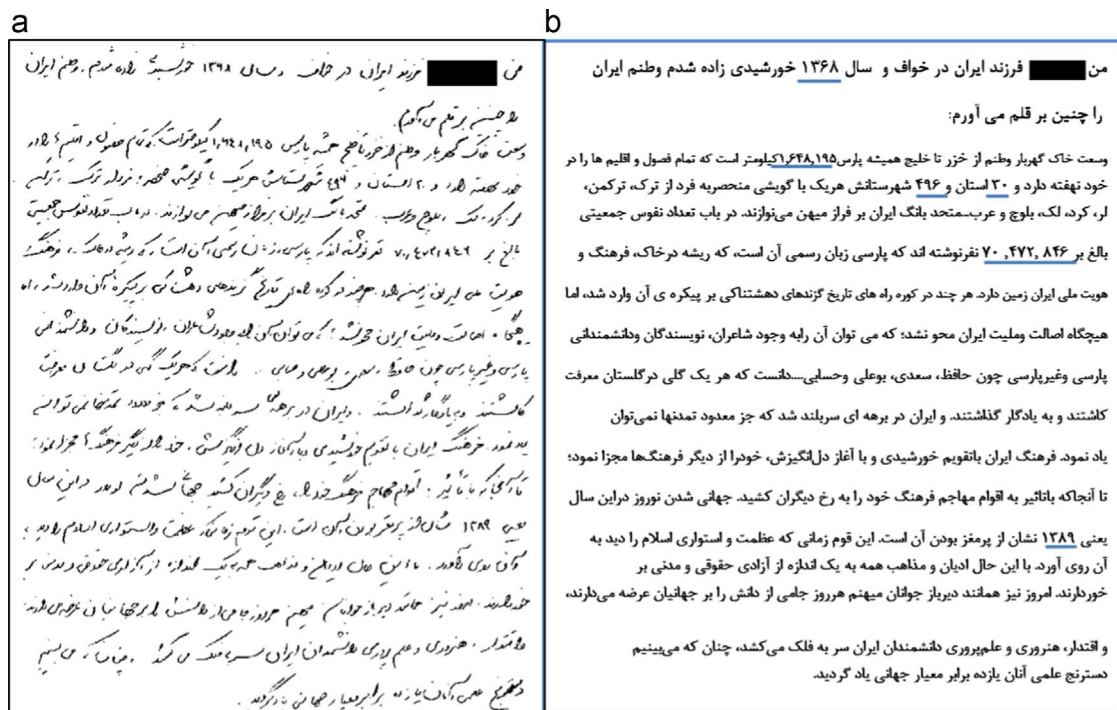


Fig. 13. A binary text sample with the corresponding ground truth content (a) handwritten sample (b) the corresponding Persian typewritten text (ground truth content). In the ground truth all the numeral strings have been underlined in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 9

Ground truth data sample: (a) an image ground truth information and (b) writer ground truth information.

Image Ground Truth		Writer Ground Truth	
Image File Name	ID0000107_P1_B43_GL.tif	Writer's ID	107
Writer's ID	107	Name & family	
Field Type	Numeral Strings	Gender	Female
Sub Field	Numeral Strings Length Five	Age	21
Name			
Content	50162	Hand-Orientation	Right-handed
Partition Type	R	Education Level	Bachelor
Image Type	Gray Level	Field of Study/ Profession	Chemist
(a)		(b)	

applications based on normalization, gray level feature extraction, and several supervised learning methods. It should be noted that providing state of the art of Persian digit recognition is not an objective of the presented example. We have conducted some recognition experiments on the grayscale digit images of this database. As shown in Table 1, 97 124 digit images were partitioned into three distinct datasets: 60 000 digit images were used for the training set, 17 124 digit images for the validation set, and 20 000 digit images for the testing set. The details of the experiment are explained in the following sections.

5.2.1. Normalization

Before extracting features, in order to have all digits the same size, an image normalization step was required. A well-known normalization technique, Aspect Ratio Adaptive Normalization (ARAN), introduced by Liu et al. (which had been applied on several scripts [27,28] also on Persian script [29]), was used in our experiments. In order to investigate the applicability of ARAN on the gray digit images, a linear forward mapping with coordinate discretization method which had shown better performance, was

applied [28]. Having conducted several experiments, the best normalized plane size was determined to be 42×42 pixels. Four normalization functions employed on the character space are shown in Eqs. (1)–(4).

$$A2 = 1 \quad (1)$$

$$A2 = A1 \quad (2)$$

$$A2 = \sqrt{\sin\left(\left(\frac{\pi}{2}\right) \times A1\right)} \quad (3)$$

$$A2 = \sqrt{A1} \quad (4)$$

In these equations, A1 and A2 refer to the aspect ratio of the digit before and after normalization, respectively (the aspect ratio is a positive value that is always less than 1). Some samples of digits before and after our normalization are depicted for digits '5' and '1' in Fig. 15.

5.2.2. Feature extraction

From among different feature extraction methods utilized for Persian handwriting recognition in the literature, gradient histogram has been outlined as a superior feature extraction method in digit recognition applications [29]. Accordingly, gradient direction features were extracted in our experiments. After many experiments, in order to extract numeral features, the image plane space was divided into 36 (6 horizontal \times 6 vertical) equal blocks; hence each block was covered by $49 (= \frac{42}{6} \times \frac{42}{6})$ pixels [30]. For each block, 4 numeral measures of gradient features were extracted as follows: for a given image (I), the magnitude (g) and direction (θ) of the gradient feature vector at each pixel (i, j) were computed as shown in Eqs. (5) and (6), respectively.

$$g(i, j) = \sqrt{g_x^2 + g_y^2} \quad (5)$$

$$\theta(i, j) = \tan^{-1} \frac{g_y(i, j)}{g_x(i, j)} \quad (6)$$

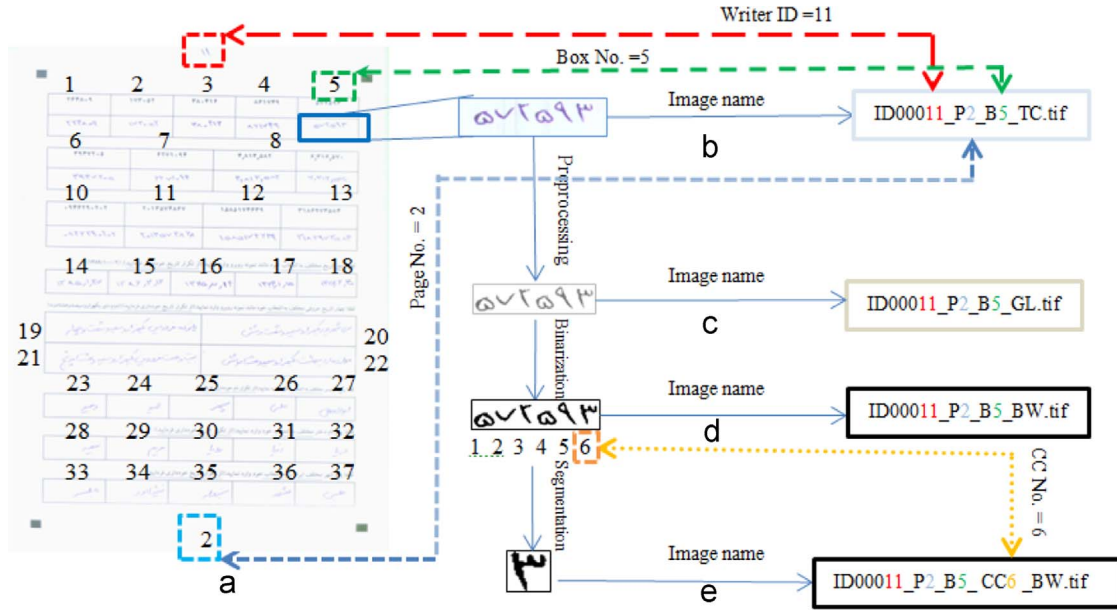


Fig. 14. Image naming process applied on all image samples of the database. (a) An HSF with numbered boxes. The box number 5 on page 2 of a participant's HSF with ID=11, is highlighted. (b–d) The true color, gray level, and binary field sample with corresponding image name, respectively, and (e) the binary segmented digit image sample with the associated filename. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

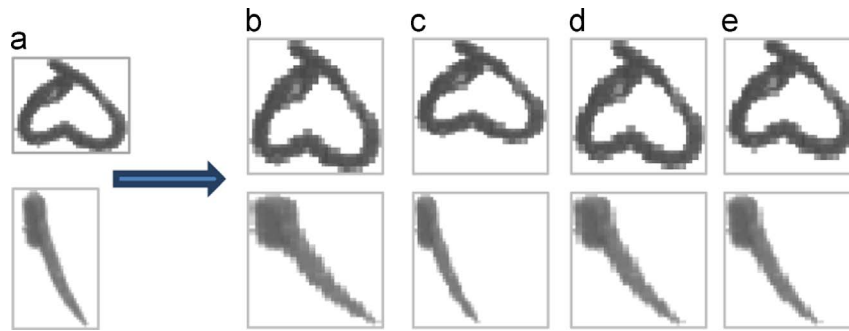


Fig. 15. Normalization effect on handwritten digit samples '5' (the upper images) and '1' (the lower images): (a) before normalization and (b–e) after applying functions corresponding to Eqs. (1)–(4), respectively.

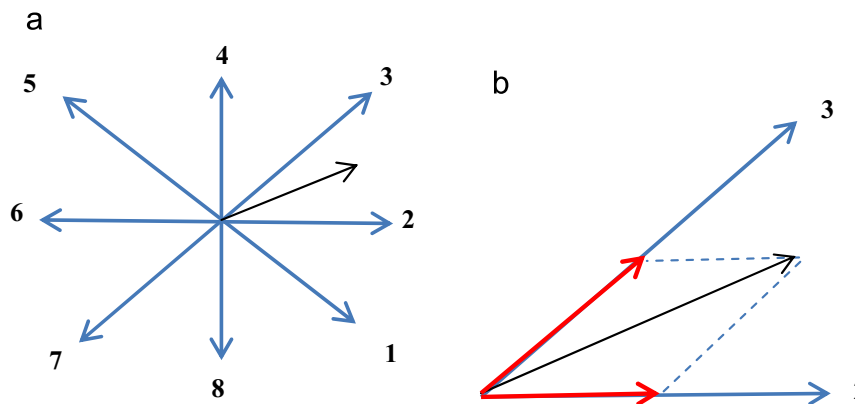


Fig. 16. Gradient features 8 main directions (4 orientations) and vector decomposition: (a) 4 main orientations and (b) vector decomposition to two red vectors on the main adjacent directions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where the gradient's vertical and horizontal components i.e., g_y and g_x , were calculated using the Sobel operator (7) [31].

$$\begin{aligned}
 g_x(i,j) &= I(i-1,j+1) + 2 \times I(i,j+1) + I(i+1,j+1) \\
 &\quad - I(i-1,j-1) - 2 \times I(i,j-1) - I(i+1,j-1) \\
 g_y(i,j) &= I(i-1,j-1) + 2 \times I(i-1,j) + I(i-1,j+1) \\
 &\quad - I(i+1,j-1) - 2 \times I(i+1,j) - I(i+1,j+1)
 \end{aligned}
 \quad (7)$$

In our experiments 8 main directions with 45° intervals were used such that opposite directions were considered as the same orientation, as seen in Fig. 16a. Any gradient vector lying between a pair of adjacent main directions was projected onto the two main adjacent directions as shown in Fig. 16b.

For each orientation in a block, the sum of magnitudes was calculated and reported as the orientation histogram feature of

that block. As a result, a 144D vector of gradient histogram features (composed of 36 blocks \times 4 orientations) for each image was constructed. Finally, in order to map all features in these vectors between 0 and 1, a simple linear normalization function was used (8).

$$F(x_{if}) = \frac{x_{if} - \min(f)}{\max(f) - \min(f)} \quad (8)$$

where x_{if} is the f th feature value of the i th sample.

5.2.3. Classification

For the classification step, we employed three distinct supervised learning methods: K Nearest Neighbor (KNN), Decision Tree (DT), and Multi-Layer Perceptron (MLP).

KNN as a lazy learner was used which finds the K closest training patterns to a testing image. In our experiments, closeness is defined in terms of Euclidean distance of our feature vectors and the parameter K in different experiments was set equal to 1, 3, 4, and 5 closest patterns. DT was used as the second classifier. DT builds a model (tree) on the training data. In order to predict the response of a testing pattern the nodes (decisions) should be followed from the root to a leaf node. In our experiments, the Classification and Regression Tree (CART) algorithm was utilized with Gini's Diversity Index (GDI) [32]. As the third classifier, an Multi-Layer Perceptron (MLP) classifier was used with three layers: input, hidden, and output layers. For training MLP, the back-propagation algorithm with the aim of minimizing the min-squared error criterion was used. The input layer had 144 neurons, which is the same number of neurons as the dimension of the input feature vector. The number of neurons in the hidden layer was chosen to be 10, 50 and 100 in different experiments and the output layer had 10 neurons, each corresponding to one digit class (0–9). A sigmoid transfer function was used for the neurons in the hidden layer and a logistic transfer function was used for the neurons in the output layer.

We conducted all our experiments on an Asus machine with 6 GB RAM and Intel[®] core[™] i5-4200 CPU@2.3 GHz using Microsoft Windows[®] 7. The programming environment was Matlab R2012b (by MathWorks Inc.). See Table 10 for the achieved recognition rate and running performance in our experiments.

As shown in Table 10, 32 distinct experiments were conducted (consider 8 classifier functions \times 4 normalization functions) through which the best result in terms of accuracy on the testing set was 98.57%, achieved by MLP when the hidden layer neurons was set to 100 and the normalization function was as Eq. (3). To make a comparison, some of well-known works preformed on Persian handwritten digit recognition have been tabulated in Table 11. Considering previous works, we achieved comparable results in terms of the number of features, viz. 144 features, and number of training and testing samples, viz. 60,000 and 20,000

instances, respectively. These experiments and their results show the applicability of our database for this kind of research and investigations on Persian handwriting recognition. However, state of the art and novel approaches of digit recognition were not the goal of the designed experiment.

6. Discussion and comparison

An overview of our database and its relevant statistics show that this database is unique among all similar scripts, such as Arabic, Dari, Urdu, and Pashto, in terms of number of words, digits, characters, symbols, texts, variety of different items, folders, and modality of all images (true color, gray level and binary), etc. Moreover, among other scripts it is a very competitive database, and in the majority of cases it is unrivaled. The uniqueness of the current work compared to similar works reported in the literature for all scripts, is shown in Table 12. Our unique database can be used for creating many useful applications and conducting important researches and studies in machine learning, pattern recognition, image processing, and document analysis, as mentioned in the previous section.

7. Conclusion and future works

The lack of a generalized, unified benchmark database for Persian handwritten script motivated us to create the first comprehensive database which alleviates most of the research demands for offline recognition of Persian handwritten texts. Furthermore, we introduced a novel comprehensive framework for creating handwriting databases for offline recognition in other scripts. HSFs were collected from both Persian males and females equally (250 per gender), and their extensive handwritten items were extracted and organized. As a result, our database includes a large number of digits, numeral strings, touching digits, worded and numeral dates, Persian alphabetical letters, common words in financial documents, punctuation marks, symbols and an unconstrained text which contains samples that captured the most common challenges of free handwritten text in Persian. Detailed ground truth files were provided for each stored image to facilitate working with this database. A unique naming strategy was created and followed for all samples and all images in the database were presented in three formats: true color, grayscale, and binary. The database has been partitioned into three separate datasets of training, validation, and testing. Our database provides a very good facility for conducting various machine learning, pattern recognition, and image processing studies as well as for creating important applications such as check recognition, mail sorting, etc. We showed an example of how to use this database for research. This

Table 10

Achieved recognition rates and running performances of 32 experiments based on the testing dataset of digits database: the best recognition rate and the corresponding running time is highlighted. Here K is the parameter of KNN and N is the number of neurons in the hidden layer of our neural network. All these results have been shown for four different normalizations.

	KNN				Decision tree	MLP			Normalization function
	$K=1$	$K=3$	$K=4$	$K=5$		$N=10$	$N=50$	$N=100$	
Recognition rate (%)	96.75	97.12	97.3	97.16	91.27	93.85	97.04	97.34	$R2=1$
Running time (s)	213.7	218.1	220.7	219.9	45.8	150.2	631.3	1080	
Recognition rate (%)	97.38	97.64	97.74	97.6	83.88	94.46	97.42	97.87	$R2=R1$
Running time (s)	218.9	223.5	221.4	222	45.5	346.5	940	1900	
Recognition rate (%)	98.04	98.2	98.29	98.1	92.24	91.88	97.82	98.22	$R2=\sqrt{R1}$
Running time (s)	217.7	219.7	220.2	220.8	46.9	226.8	687.8	1.35	
Recognition rate (%)	98.09	98.24	98.34	98.21	92.33	94.07	97.97	98.57	$R2=\sqrt{\sin((\pi/2) \times R1)}$
Running time (s)	217.6	214.3	223.5	219.7	48.2	364	647.8	1433	

Table 11

A comparison on results of the state of the art works on Persian handwritten digit recognition.

Ref.	Method	No. of features	Classifier	Dataset	No. of training samples	No. of testing samples	Accuracy on test set (%)
[33]	Wavelet transform	64	SVM	Personal	2240	1600	92.44
[34]	Asymmetrical segmentation	12	NN ^a	Personal	230	500	97.6
[29]	Gradient histogram	200	CFPC ^b	CENPARMI [7]	11 000	2000	99.16
[35]	Template matching	60	NN	Personal	6000	4000	97.65
[7]	Profile	64	SVM	CENPARMI [7]	11 000	5000	97.32
[36]	Profile-cross count	257	SVM	personal	3939	4974	99.57
[37]	Chain-code	196	SVM	HODA [21]	60,000	20,000	99.02
Current research	Gradient histogram	144	MLP	Current DB	60 000	20 000	98. 57

^a Neural network.^b Class-specific feature polynomial classifier.**Table 12**

A comparison of our database with state of the art databases throughout the handwriting recognition community.

Database name	Language/ script	No. of writers	Worded dates	Numeral dates	Digits	Numeral strings	Touching digits	Alphabet letters	Symbols	Words	Texts/ paragraphs
CEDAR [11]	English	N/A	–	–	21,179	–	–	27,837	–	10,570	–
IAM [38]	English	400	–	–	–	–	–	–	–	82,227	–
GRUHD [39]	Greek	1000	–	–	123 256	–	–	–	–	102,692	–
Ref. [40]	Italian	277	–	–	28,678	–	–	66,609	–	48,584	–
Hit-Mw [14]	Chinese	780	–	–	–	–	–	186,44	–	–	–
Ref. [4]	Arabic	N/A	–	–	15,175	–	–	–	–	4998	–
CENPARMI [41]	Arabic	328	–	284	46 800	13 439	–	21 426	1640	11,375	–
AHTID/MW [42]	Arabic	53	–	–	–	–	–	–	–	22,896	–
KHATT [19]	Arabic	1000	–	–	–	–	–	–	–	–	6000
CENPARMI [43]	Dari	200	–	200	28,000	7600	–	7400	1400	14,600	–
CENPARMI [7]	Persian	175	–	175	18,000	7350	–	11,900	–	8575	–
FHT [20]	Persian	250 (163+92) Men and women	–	–	–	–	–	–	–	–	1000
IFHCDB [44]	Persian	N/A	–	–	17,740	–	–	52,380	–	–	–
HODA [21]	Persian	N/A	–	–	102 357	–	–	–	–	–	–
IAUT/PHCN [45]	Persian	380	–	–	–	–	–	–	–	34,200	–
IFN/Farsi-data-base [46]	Persian	600 (340+260) Men and Women	–	–	–	–	–	–	–	7271	–
CENPARMI [47]	Persian	400	295	–	24,121	13,334	–	–	2738	516	–
Our database	Persian	500 (250+250) Men and women	2000	2500	97 124	19,500	460	42,500	16,000	70,000	500

database is available in its entirety to the research community. In the future, we are going to extend our database and its collections by providing a large number of samples of Persian historical documents.

Conflict of interest

None declared.

Acknowledgment

The names and family names of all the 500 participating writers of this database have been removed from all images and ground truth files in our database for confidentiality reasons. All our writers have been represented by 3-digit IDs. The authors of this paper wish to thank all 500 anonymous writers for their contribution and their time spent completing the 7 pages of the HSFs.

References

- [1] R. Plamondon, S.N. Srihari, Online and off-line handwriting recognition: a comprehensive survey, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (2000) 63–84.
- [2] F. Kimura, M. Shridhar, Handwritten numerical recognition based on multiple algorithms, *Pattern Recognit.* 24 (1991) 969–983.
- [3] H.-J. Winkler, M. Lang, Online symbol segmentation and recognition in handwritten mathematical expressions, In: *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997, ICASSP-97, IEEE, 1997*, pp. 3377–3380.
- [4] Y. Al-Ouali, M. Chérret, C. Suen, *Databases for recognition of handwritten Arabic cheques*, *Pattern Recognit.* 36 (2003) 111–121.
- [5] L.M. Lorigo, V. Govindaraju, Offline Arabic handwriting recognition: a survey, *IEEE Trans. Pattern Anal. Mach. Intell.* 28 (2006) 712–724.
- [6] I. Guyon, R.M. Haralick, J.J. Hull, I.T. Phillips, Data sets for OCR and document image understanding research, *Handbook of Character Recognition and Document Image Analysis, 1997*, 779–799.
- [7] F. Solimanpour, J. Sadri, C.Y. Suen, Standard databases for recognition of handwritten digits, numerical strings, legal amounts, letters and dates in Farsi language, In: *Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition La Baule, France, 2006*, pp. 3–7.
- [8] S. Izadi, J. Sadri, F. Solimanpour, C.Y. Suen, *A Review on Persian Script and Recognition techniques, Arabic and Chinese Handwriting Recognition, Springer (2008)*, p. 22–35.
- [9] S. Khalighi, P. Tirdad, H.R. Rabiee, A novel OCR system for calculating handwritten Persian arithmetic expressions, In: *Proceedings of the 2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), IEEE, 2009*, pp. 277–282.
- [10] J. Sadri, Y. Akbari, M.J. Jalili, A. Farahi, M. Habibi, A new system for recognition of handwritten Persian bank checks, In: *Proceedings of the 2011 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2011*, pp. 925–930.
- [11] J.J. Hull, A database for handwritten text recognition research, *IEEE Trans. Pattern Anal. Mach. Intell.* 16 (1994) 550–554.
- [12] P.J. Grother, NIST special database 19 handprinted forms and characters database, *Natl. Inst. Stand. Technol.* (1995).
- [13] U.-V. Marti, H. Bunke, The IAM-database: an English sentence database for offline handwriting recognition, *Int. J. Doc. Anal. Recognit.* 5 (2002) 39–46.
- [14] T. Su, T. Zhang, D. Guan, Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text, *Int. J. Doc. Anal. Recognit.* 10 (2007) 27–38.

- [15] K. Dae-Hwan, Y.-S. HWANG, P. Sang-Tae, K. Eun-Jung, P. Sang-Hoon, B. Sung-Yang, Handwritten Korean character image database PE92, *IEICE Trans. Inf. Syst.* 79 (1996) 943–950.
- [16] M. Nakagawa, T. Higashiyama, Y. Yamanaka, S.-i. Sawada, L. Higashigawa, K. Akiyama, On-line handwritten character pattern database sampled in a sequence of sentences without any writing instructions, In: *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, 1997, IEEE, 1997, pp. 376–381.
- [17] U. Bhattacharya, B. Chaudhuri, Databases for research on recognition of handwritten characters of Indian scripts, In: *Proceedings of the 2005 Eighth International Conference on Document Analysis and Recognition*, IEEE, 2005, pp. 789–793.
- [18] M. Pechwitz, S.S. Maddouri, V. Märgner, N. Ellouze, H. Amiri, IFN/ENIT – database of handwritten Arabic words, *Colloque international francophone sur l'écrit et le document, Hammamet, Tunisia* (2002), p. 129–136.
- [19] S.A. Mahmoud, I. Ahmad, M. Alshayeb, W.G. Al-Khatib, M.T. Parvez, G.A. Fink, V. Margner, H.E. Abed, KHATT: Arabic offline handwritten text database, In: *Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition*, IEEE Computer Society 2012, pp. 449–454.
- [20] M. Ziaratban, K. Faez, F. Bagheri, F.H.T., An unconstrained Farsi handwritten text database, in: *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ICDAR'09. IEEE, 2009, pp. 281–285.
- [21] H. Khosravi, E. Kabir, Introducing a very large dataset of handwritten Farsi digits and a study on their varieties, *Pattern Recognit. Lett.* 28 (2007) 1133–1141.
- [22] N. Otsu, A thresholding selection method from graylevel histogram, *IEEE Trans. Syst. Man Cybern.* 9 (1979) 62–66.
- [23] J.S. Lim, Two-Dimensional Signal and Image Processing, Englewood Cliffs, 1990.
- [24] J. Sadri, C.Y. Suen, T.D. Bui, A genetic framework using contextual knowledge for segmentation and recognition of handwritten numeral strings, *Pattern Recognit.* 40 (2007) 898–919.
- [25] Ad.S. Britto Jr., R. Sabourin, F. Bortolozzi, C.Y. Suen, The recognition of handwritten numeral strings using a two-stage HMM-based method, *Int. J. Doc. Anal. Recognit.* 5 (2003) 102–117.
- [26] A. Gillies, E. Erlanson, J. Trenkle, S. Schlosser, Arabic text recognition system, In: *Proceedings of the Symposium on Document Image Understanding Technology*, 1999.
- [27] C.-L. Liu, K. Nakashima, H. Sako, H. Fujisawa, Handwritten digit recognition: benchmarking of state-of-the-art techniques, *Pattern Recognit.* 36 (2003) 2271–2285.
- [28] C.-L. Liu, K. Nakashima, H. Sako, H. Fujisawa, Handwritten digit recognition: investigation of normalization and feature extraction techniques, *Pattern Recognit.* 37 (2004) 265–279.
- [29] C.-L. Liu, C.Y. Suen, A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters, *Pattern Recognit.* 42 (2009) 3287–3295.
- [30] M. Cheriet, N. Kharma, C.-L. Liu, C. Suen, *Character Recognition Systems: A Guide for Students and Practitioners*, John Wiley & Sons, 2007.
- [31] G. Srikanthan, S.W. Lam, S.N. Srihari, Gradient-based contour encoding for character recognition, *Pattern Recognit.* 29 (1996) 1147–1160.
- [32] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, *Classification and Regression Trees*, CRC Press, 1984.
- [33] H.M.M. Hosseini, A. Bouzerdoum, A combined method for Persian and Arabic handwritten digit recognition, In: *Proceedings of the 1996 Australian and New Zealand Conference on Intelligent Information Systems*, IEEE, 1996, pp. 80–83.
- [34] A. Harifi, A. Aghagolzadeh, A. New Pattern For Handwritten Persian/Arabic digit recognition, in: *Proceedings of the International Conference on Information Technology (ICT2004)*, Istanbul, 2005, pp. 130–133.
- [35] M. Ziaratban, K. Faez, F. Faradji, Language-based feature extraction using template-matching in Farsi/Arabic handwritten numeral recognition, In: *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, IEEE, Curitiba, Parana, Brazil 2007, pp. 297–301.
- [36] H. Soltanzadeh, M. Rahmati, Recognition of Persian handwritten digits using image profiles of multiple orientations, *Pattern Recognit. Lett.* 25 (2004) 1569–1576.
- [37] A. Alaei, P. Nagabhushan, U. Pal, Fine classification of unconstrained handwritten Persian/Arabic numerals by removing confusion amongst similar classes, In: *Proceedings of the 10th International Conference on Document Analysis and Recognition*, IEEE, Barcelona 2009, pp. 601–605.
- [38] U.V. Marti, H. Bunke, The IAM-database: an English sentence database for offline handwriting recognition, The IAM-database: an English sentence database for offline handwriting recognition (IJDA), 5 (2002) 39–46.
- [39] E. Kavallieratou, N. Liolios, E. Koutsogeorgos, N. Fakotakis, G. Kokkinakis, The GRUHD database of Greek unconstrained handwriting, In: *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR)*, Seattle, Washington, 2001, pp. 561–565.
- [40] G. Dimauro, S. Impedovo, R. Modugno, G. Pirlo, A new database for research on bank-check processing, In: *Proceedings of the 2002 Eighth International Workshop on Frontiers in Handwriting Recognition*, IEEE, 2002, pp. 524–528.
- [41] H. Alamri, J. Sadri, C.Y. Suen, N. Nobile, A novel comprehensive database for Arabic off-line handwriting recognition, in: *Proceedings of 11th International Conference on Frontiers in Handwriting Recognition*, ICFHR, 2008, pp. 664–669.
- [42] A. Mezghani, S. Kanoun, M. Khemakhem, H.E. Abed, A database for arabic handwritten text image recognition and writer identification, In: *Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition*, IEEE Computer Society, 2012, pp. 399–402.
- [43] M.I. Shah, J. Sadri, C.Y. Suen, N. Nobile, A. New multipurpose comprehensive database for handwritten dari recognition, In: *Proceedings of the 11th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Montréal, Québec, 2008, pp. 635–640.
- [44] S. Mozaffari, K. Faez, F. Faradji, M. Ziaratban, S. Mohamad, A. comprehensive isolated Farsi/Arabic character database for handwritten OCR research, In: *Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006.
- [45] A. Bidgoli, M. Sarhadi, IAUT/PHCN: Islamic Azad University of Tehran/Persian handwritten city names, a very large database of handwritten Persian word, In: *Proceedings of 11th International Conference on Frontiers in Handwriting Recognition (ICFHR 11)*, Montreal, Canada, 2008, pp. 192–197.
- [46] S. Mozaffari, H. El Abed, V. Märgner, K. Faez, A. Amirshahi, I. Semnan, IFN/Farsi-database: a database of farsi handwritten city names, In: *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, 2008.
- [47] P.J. Haghighi, N. Nobile, C.L. He, C.Y. Suen, A New Large-Scale Multi-purpose Handwritten Farsi database, *Image Analysis and Recognition*, Springer (2009), p. 278–286.

Javad Sadri is an assistant professor at the Computer Science and Software Engineering Department of Concordia University in Canada since 2015. He was a visiting professor at McGill center for Bioinformatics, McGill University, Montreal, Quebec, Canada in 2013–2014, and he worked as an assistant professor and chair of the department in the Computer Engineering Department of University of Birjand in Iran from 2010 to 2013. He obtained his Ph.D. (in 2007) in the field of Pattern Recognition and Machine Learning at CENPARMI (Center for Pattern Recognition & Machine Intelligence) at Concordia University, Canada. Then he spent one year of his postdoctoral research at CENPARMI and two years on bioinformatics at McGill University. He has published several papers in international journals and conferences in the areas of pattern recognition, document recognition, image processing and bioinformatics. He was one of the recipients of the best poster/paper award at the Ninth IWFHR (International Workshop on Frontiers in Handwriting Recognition) in Tokyo, Japan, in 2004. He is also one of the contributors of the book "Character Recognition Systems, a Guide for Students & Practitioners", Published in 2007, by John Wiley & Sons Inc. He was the organizer and general chair of the First Iranian Conference on Pattern Recognition and Image Analysis (PRIA2013) held in March 2013. Sadri is also a reviewer for several international journals in the field of pattern recognition, bioinformatics, and document analysis such as *Pattern Recognition*, *Pattern Recognition Letters*, *Int. J. Pattern Recognition and Artificial Intelligence*, *Int. J. Document Analysis and Recognition*, *Pattern Analysis and Applications* and *Signal Processing*. His current research interests are: applications of pattern recognition and machine learning techniques in handwritten recognition, document analysis, segmentation and recognition of historical documents, image processing, bioinformatics, affective computing, and also evolutionary optimization.

Mohammad Reza Yeganehzad received his B.Sc. degree from University of Birjand, Birjand, Iran, in 2011 and received his M.Sc. degree in IT Engineering from Urmia University of Technology, Urmia, Iran, in 2013. He is now working as network administrator at Communications Regulatory Authority (CRA) of Iran, Mashhad, Iran. His research interests include creating handwritten database, Persian handwritten features and recognition algorithms, web and network security, and web mining.

Javad Saghi received his B.Sc. degree from University of Birjand in the year 2011 and received his M.Sc. degree in Artificial Intelligence from Graduate University of Advanced Technology, Iran, in the year 2014. His research interests include: AI, pattern recognition, creating handwritten database, Persian handwritten recognition, applied computer programming.