Pattern Recognition 60 (2016) 378–393

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

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
Numeral 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.

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).

http://dx.doi.org/10.1016/j.patcog.2016.03.024
0031-3203/

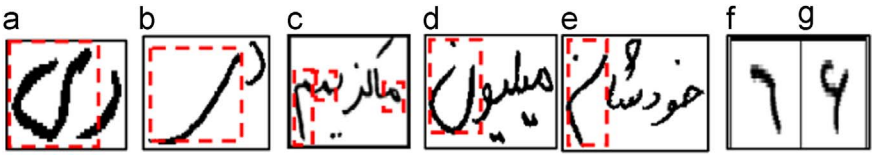

Fig. 1. Some challenges of Persian handwriting recognition: (a and b) cursiveness of the character "ی" in the word "دی" ("Dei"), (c) different shapes of the letter "م" in the word "ماکزیمم" ("maximum"), (d), formal writing style of the character "ن" written at the end of the word "میلیون" ("million"), (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 "ماکزیمم" ("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 CENPARMI[1] [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 **Pa**ttern **R**ecognition and **M**achine **I**ntelligence.