Noname manuscript No.
(will be inserted by the editor)

HKR For Handwritten Kazakh & Russian Database

Daniyar Nurseitov [1,2] · Kairat Bostanbekov [1,2] · Daniyar Kurmankhojayev [3] · Anel Alimova [1,2] · Abdelrahman Abdallah [1,2]

Received: date / Accepted: date

Abstract In this paper, we present a new Russian and Kazakh database (with about 95% of Russian and 5% of Kazakh words/sentences respectively) for offline handwriting recognition. A few pre-processing and segmentation procedures have been developed together with the database. The database is written in Cyrillic and shares the same 33 characters. Besides these characters, the Kazakh alphabet also contains 9 additional specific characters. This dataset is a collection of forms. The sources of all the forms in the datasets were generated by LaTeXwhich subsequently was filled out by persons with their handwriting. The database consists of more than 1400 filled forms. There are approximately 63000 sentences, more than 715699 symbols produced by approximately 200 different writers. It can serve researchers in the field of handwriting recognition tasks by using deep and machine learning.

Keywords Handwriting recognition · Cyrillic dataset · Kazakh · Russian

This work was done with the support of grant funding for scientific projects of the MES RK No AR05135175 Development and implementation of a system for recognizing handwritten addresses of written correspondence JSC KazPost using machine learning.

Daniyar Nurseitov
nurseitovdb@gmail.com

Kairat Bostanbekov
kairat.boss@gmail.com

Daniyar Kurmankhojayev
kurman.daniyar@gmail.com

Anel Alimova
anic2002@mail.ru

Abdelrahman Abdallah
MSc Machine Learning & Data Science
Satbayev University
abdoelsayed2016@gmail.com

[1] Satbayev University Almaty, Kazakhstan [2] National Open Research Laboratory for Information and Space Technologies, Almaty, Kazakhstan [3] Hong Kong Polytechnic University, Hung Hom, Hong Kong

1 Introduction

Today, handwriting recognition is a very urgent task. A solution to this problem would automate the business processes of many companies. One of the clear examples is a postal company, where the task of sorting a large volume of letters and parcels is an acute issue. Many researchers have made different types of handwritten text recognition systems for different languages such as English [2,3,4], Chinese [5], Arabic [9], Japanese [6], Bangla [7], Malyalam [8], etc. Having said that, the recognition problems of these scripts cannot be considered be entirely solved.

Any language contains a large number of words. For example, dictionaries of the Russian and Kazakh languages on average register more than 100,000 words, and the Oxford English dictionary more than 300,000 words. In this regard, collecting an exhaustive database of handwritten words, which include all words with a large variation in handwriting, seems almost impossible. In other words, there is always a word that the system cannot recognize. To the best of our knowledge, the analogs of handwritten text database for Russian and Kazakh languages do not exist.To create such a database, we decided to adopt the general principles of data collection and storage described in the IAM Database [1]. In the context of handwritten address recognition, it is necessary to identify the many keywords that can occur in the address. We utilized three different datasets described as following:

- Handwritten samples (Forms) of keywords in Kazakh and Russian (Areas, Cities , Village , etc.)
- Handwritten Kazakh and Russian alphabet in cyrillic
- Handwritten samples (Forms) of poems in Russian

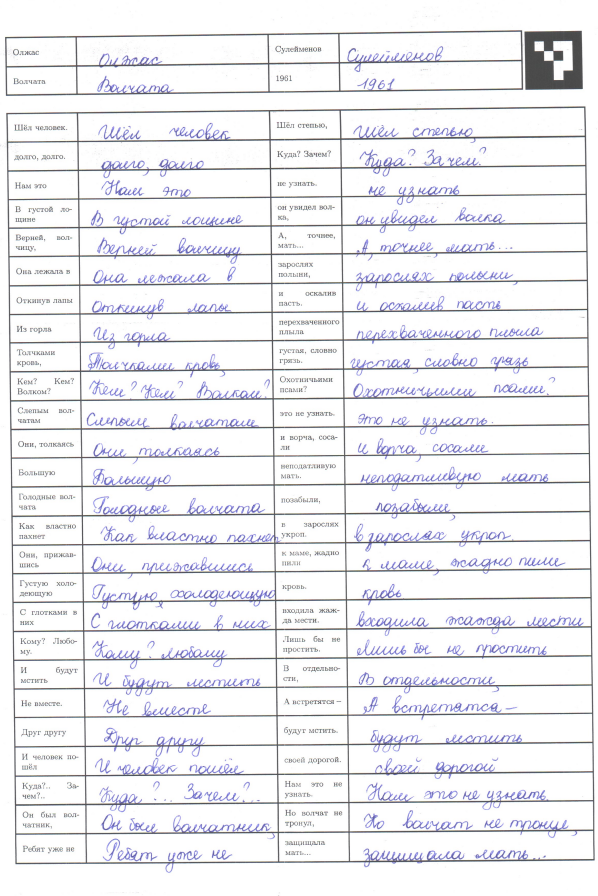

Fig. 1 One of the poem form in the dataset

In this paper, we describe the first version of a database that contains Russian words and also present a new database for offline handwriting recognition. The collection of this database combines the following steps. As an initial step, we collected the first data set with our own hands, since it is almost impossible to find such a set publicly available. This dataset was obtained by using forms, which consisted of machine-typed texts, and empty lines next to those texts. These empty lines were subsequently filled out by persons with their handwriting. It can serve as a basis for a variety of handwriting recognition tasks. Second, we collected handwritten Kazakh and Russian alphabet in Cyrillic. The last set of data came from handwritten samples of poems also filled by our own hands in Russian language.The databases were produced by approximately 200 different writers, each having 5 to 10 forms (made up of poem and keyword texts) to fill.

For these purposes, we determined the minimum set of words, which includes all the names of cities, towns, villages, districts, and streets in Kazakhstan, and created layouts for filling out forms. Forms were created in such a way as to simplify the process of cutting words from the form as much as possible (Fig. 1). Extensive experiments related to the preprocessing of forms were also carried out in order to automatically identify forms, determine the contours of forms, compensate rotations, and also remove edge artifacts at the boundaries of segmented words.

To solve the problem of recognition and processing of natural languages (natural language processing), which consists of optical recognition of characters of the manuscript texts in Russian and Kazakh languages, innovative software is being developed using state-of-the-art neural network-based machine learning methods.

The following section defines the related work on Handwriting Databases. Section 3 presents Data collection and storage phases is one of the most time consuming and costly stages. Section 4 provides Automated Labeling and Words Segmentation. Section 5 provides further characteristics of the Database, and concluding and future work are given in Section 6.

2 Related Work

The IAM Handwriting Database [1,10] comprises handwritten samples in English which can be used to evaluate systems like text segmentation, handwriting recognition, writer identification and writer verification. The database is developed on the Lancaster-Oslo/Bergen Corpus and comprises forms where the contributors copied a given text in their natural unconstrained handwriting. Each form was subsequently scanned at 300 dpi and saved as gray level (8-bit) PNG image.The IAM Handwriting Database 3.0 includes contributions from 657 writers, making a total of 1539 handwritten pages comprising 5685 sentences, 13,353 text lines, and 115,320 words. The database is labeled at the sentence, line, and word levels. It has been widely used in word spotting [11, 12,13,14], writer identification [15,16,17,18,19], handwritten text segmentation [20,21,22]and offline handwriting recognition [23,24,25,26].

RIMES [27] is a representative database of an industrial application. The main idea of developing this database was to collect handwritten samples similar to those that are sent to different companies by postal mail and fax by individuals. Each contributor was assigned a fictitious identity and a maximum of up to five different scenarios from a set of nine themes. These themes included real-world scenarios like damage declaration or modification of contract. The subjects were required to compose a letter for a given scenario using their own words and layout on a white paper using black ink. A total of 1300 volunteers contributed to data collection, providing 12,723 pages corresponding to 5605 mails. Each mail contains two to three pages, including the letter written by the contributor, a form with information about the letter, and an optional fax sheet. The pages were scanned, and the complete database was annotated to support evaluation of tasks like document layout analysis [28], mail classification [29], handwriting recognition [30] and writer recognition [17]. The National Institute of Standards and Technology, NIST, developed a series of databases [31] of handwritten characters and digits supporting tasks like isolation of fields, detection and removal of boxes in forms, character segmentation, and recognition.The form comprises boxes containing writer information, 28 boxes for numbers and 2 for alphabets, and 1 box for a paragraph of text. The NIST Special Database 1 comprised samples contributed by 2100 writers. The latest version of the database, the Special Database 19, comprises handwritten forms of 3600 writers with 810,000 isolated character images along with ground truth information. This database has been widely employed in a variety of handwritten digits [32] and character recognition systems [33].

CVL [34] is a database of handwritten samples supporting handwriting recognition, word spotting, and writer recognition. The database consists of seven different handwritten

texts, one in German and six in English. A total of 310 volunteers contributed to data collection, with 27 authors producing 7 and 283 writers providing 5 pages each. The ground truth data is available in XML format, which includes transcription of text, the bounding box of each word, and the identity of the writer. The database has been used for writer recognition and retrieval [35] and can also be employed for other recognition tasks. In addition to regular text, a database of handwritten digit strings written by 303 students has also been compiled [36]. Each writer provided 26 different digit strings of different lengths, making a total of 7800 samples. Isolated digits were extracted from the database to form a separate datasetthe CVL Single Digit Dataset. The Single Digit Dataset comprises 3578 samples for each of the digit class (0-9). A subset of this database has also been used in the ICDAR 2013 digit recognition competition [36].

The AHDB [37,38] is an offline database of Arabic handwriting together with several pre-processing procedures. It contains Arabic handwritten paragraphs and words. Words used to represent numbers on checks produced by 100 different writers. The database was mainly intended to support automatic processing of bank checks, but it also contains pages of unconstrained texts allowing evaluation of generic Arabic handwriting recognition systems as well. The database was employed in handwriting recognition [39] and writer identification tasks [40].

3 Data collection and storage

3.1 Data collection

A data collection phase is one of the most time consuming and costly stages. Our main task is to simplify and automate as much as possible. The sources of all the forms in the datasets were generated by LaTeX, then converted to PDF and printed to be filled by writers. So, it was an easy task to generate the correct labels for the printed text on the forms.Each writer filled approximately between 5-10 forms from keyword and poem forms,so each form in the dataset is written approximately between 5-10 writers.Each form has a unique id at the name of the form.. The word or letter is placed in the rectangle.The filled forms and letters were been scanned with a Canon MF4400 Series UFRII scanner at a resolution of 300 dpi and a color depth of 24 bits.

We collected three different Datasets described as the following:

- Handwritten samples (Forms) of keywords in Kazakh and Russian (Areas, Cities , Village , etc.) are shown in Fig. 2.
- Handwritten Kazakh and Russian alphabet in Cyrillic are shown in Fig. 2.
- Handwritten samples (Forms) of poems in Russian are shown in Fig. 1.

The next step was to annotate the collected data, i.e. to mark and synthesize new samples from existing ones, using various geometric and photo-metric transformations (data augmentation).

3.1.1 Keyword Database

To begin with, we consider correspondence addresses relevant for the Republic of Kazakhstan, as the list of keywords containing the following names:

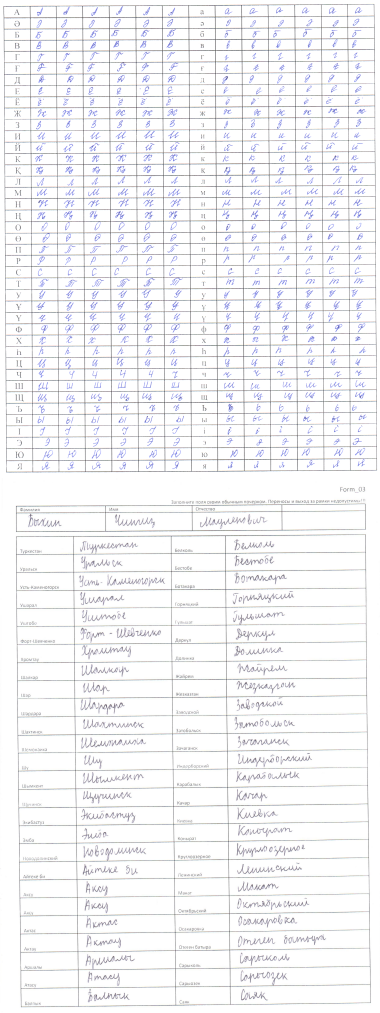

Fig. 2 Two forms for collecting handwritten samples of the Cyrillic alphabet and Keywords.

- Areas
- Cities
- Village
- Settlements
- Streets
- Poems
- Russian Letter

Additional information, such as:

- Indices
- Phones
- Surnames
- Company Names

were not included in the database.

3.1.2 Handwritten Alphabet and Forms

There are two fundamental approaches to text recognition: character recognition (Optical Character Recognition, OCR) and word recognition (Optical Word Recognition, OWR). With OCR, a model dataset required to train the model should contain handwritten samples of all the characters in a language alphabet. It is an important for each language to

compose separate forms, since the set of letters of different alphabets can vary greatly. On the other hand, with OWR, a model dataset required to train the model should contain handwritten samples of all the Words for the language. Further, for subsequent training and testing of the model, handwritten samples of target words are needed. An example of one of the forms for collecting word samples and letters (Fig. 2).

3.1.3 Data Collection Methods

A person who has agreed to provide a sample of his handwriting will fill the forms and give the form to us and we scan and save it in our database.

4 Automated Labeling and Words Segmentation

4.1 Automated Labeling

Labeled data are data that have been marked with labels identifying certain features, characteristics, or a kind of object. The labeling of data is a prerequisite for recognition experiments. Labeling data is expensive, time consuming, and error prone."like in IAM Dataset" [1], we decided to do as much automation as possible automatically. The sources of all the forms printed (and subsequently filled by writers) were saved on a text file with a unique id for the form and the cell number in the form. So it was an easy task to generate the correct labels for the printed text on the forms. In this regard, we have developed a recommendation system that allows us to simplify the process of labeling data in forms.

4.2 Segmentation

The form is designed so that it is possible to easily identify and segment by cells. To identify the form, there is a marker in the upper-right corner of each form. To simplify the process of segmentation, the entire form is divided by horizontal and vertical lines, which makes it quite easy to restore the structure of the document, and accordingly, the spatial position of the word. Words are indexed (annotated) according to their position in the table. In order to cut out cells from the form, the following actions (pre-processing) are performed:

- filtering forms to enhance table boundaries
- defining of the contours of the table
- Determination and compensation of the angle of rotation
- exclusion of lines
- Sorting forms by id (marker)
- Streaming the division of forms into words
- name and storage of words

After the image areas corresponding to the word cells are segmented, segmented image areas corresponding to the word cells may contain some edge artifacts. For example, line artifacts cut out with a cell or parts of a word from a neighboring cell can be attributed to artifacts (Fig. 3). We eliminate these artifacts by constructing vertical and horizontal histograms (Fig. 4 , Fig. 5) also by cutting off parts that are separately localized closer to the edges of the cell.

However, it is not always possible to eliminate all artifacts. The following are some aspects that make further processing of a segmented word difficult:

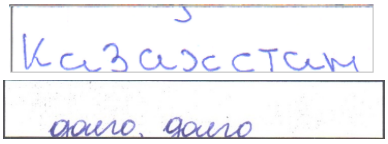

Fig. 3 Example of a region cut out of a form with a word. A pronounced cell line and a piece of letter from a neighboring area are visible along the edges.

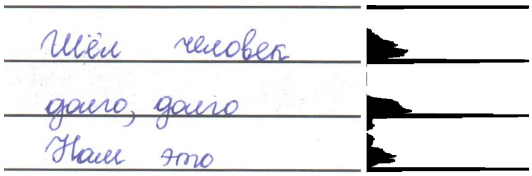

Fig. 4 Horizontal histogram

Fig. 5 Vertical histogram

- Letters may not be interconnected.
- Letters can be perfected with artifacts.
- The position of the letters and their size vary significantly from word to word.
- Letters can be written in different colors (blue, black, red).

In this regard, we have developed a recommendation system that allows us to simplify the process of selecting areas with words from the form.

- We suggest filling out the form using the blue pen. This will allow the system to distinguish the word from the table borders at the color level. For example, by converting an image from RGB to HSV, we get a color representation of objects that is invariant with respect to lighting. In this color space, blue remains blue, regardless of the brightness and intensity of the image.
- sometimes eliminating parts of words from neighboring cells is impossible without distorting the target content of a given cell; therefore, when filling out the form, it is desirable that the subject does not go beyond the boundaries of the cell.
- find the region of interest (ROI) in forms ,the ROI in our forms is two columns that are filled by writers.
- We segmented the cells depending on the horizontal white space between the cell by using the histogram (Fig. 4).
- then we exclusion of lines that cross the words.
- Finally, we cropped and segmented the cells depending on the vertical white-space by using the histogram (Fig. 5).

Next, we normalize the words images, as follow :

- cast all word images to the same size
- center the word on the image
- reduce image size

The final image shown in Fig. 6.

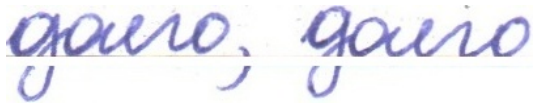

Fig. 6 Examples of segmented words.

5 Further Characteristics of the Database

The database consists of more than 1400 filled forms written by 200 writers. There are approximately 63000 sentences, more than 715699 symbols shown in Fig. 7. And also There are approximately 106718 words. total images in the dataset after pre-processing and segmentation the forms are 64943 images.

6 Conclusion and Future Work

We have built the handwritten Kazakh, Russian database. The database can serve as a basis for research in handwriting recognition. This contains Russian Words (Areas, Cities, Village, Settlements, Areas, Streets) by a hundred different writers. It also incorporates the most popular words in the Republic of Kazakhstan.A few pre-processing and segmentation procedures have been developed together with the database. Finally, it contains free handwriting forms in any area of the writer interest. This database is meant to provide a training and testing set for Kazakh, Russian Words recognition research.In future, as further work on gathering Handwriting samples of keywords and envelope shots will continue.At the same time, envelopes are annotated and various metrics are checked to evaluate the recognition error.In order for the artifacts to not interfere, we need to collect as much tagged data as possible.

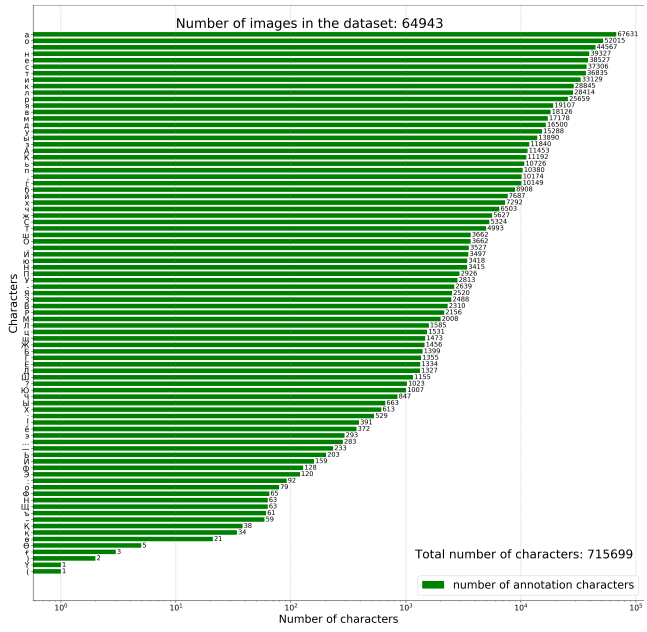

Fig. 7 Histogram of Characters in the dataset

Acknowledgements This work was funded by the Ministry of Education and Science of the Republic of Kazakhstan (Grant No AP05135175)

References

1. U. Marti and H. Bunke, A full English sentence database for off-line handwriting recognition, In Proc. of the 5th Int. Conf. on Document Analysis and Recognition, pages 705 - 708, (1999).
2. H. Liu and X. Ding, Handwritten Character Recognition using Gradient Feature and Quadratic Classifier with Multiple Discrimination Schemes, Proc. 8th Int. Conf. on Document Analysis and Recognition, pp. 19-25, (2005).
3. Fischer, Andreas, Ching Y. Suen, Volkmar Frinken, Kaspar Riesen, and Horst Bunke, A fast matching algorithm for graph-based handwriting recognition.In Graph-Based Representations in Pattern Recognition, pp. 194-203. Springer Berlin Heidelberg, (2013).
4. Zamora-Martnez, Francisco, Volkmar Frinken, Salvador Espaa-Boquera, Maria Jose Castro-Bleda, Andreas Fischer, and Horst Bunke, Neural network language models for off-line handwriting recognition, Pattern Recognition 47, no. 4 (2014): 1642-1652.
5. Tao, Dapeng, Lingyu Liang, Lianwen Jin, and Yan Gao, Similar handwritten Chinese character recognition by kernel discriminative locality alignment, Pattern Recognition Letters 35 (2014): 186-194.
6. Das, Soumendu, and Sreeparna Banerjee, An Algorithm for Japanese Character Recognition, International Journal of Image, Graphics and Signal Processing (IJIGSP) 7, no. 1 (2014): 9.
7. U. Bhattacharya, M. Shridhar, S. K. Parui, P. K, Sen and B. B. Chaudhuri, Offline recognition of ine recognition of handwritten Bangla characters: an efficient two-stage approach, Pattern Analysis and Applications, vo1.l5(4), pp.445-458, (2012).
8. John, Jomy, and Kannan Balakrishnan, A system for offline recognition of handwritten characters in Malayalam script, International Journal of Image, Graphics and Signal Processing (IJIGSP) 5, no. 4 (2013):53.
9. Parvez, Mohammad Tanvir, and Sabri A. Mahmoud, Arabic handwriting recognition using structural and syntactic pattern attributes, Pattern Recognition 46, no. 1 (2013): 141-154.
10. UV Marti, H Bunke, The iam-database: An english sentence database for offline handwriting recognition, Int. J. Doc. Anal. Recognit. 5(1), 3946 (2002).
11. S Wshah, G Kumar, V Govindaraju, in Proceedings of International Conference on Frontiers in Handwriting Recognition, Script independent word spotting in offline handwritten documents based on hidden markov models, pp. 1419, (2012).
12. S Wshah, G Kumar, V Govindaraju, in Proceedings of 21st International Conference on Pattern Recognition, Multilingual word spotting in offline handwritten documents,pp. 310313,(2012).
13. A Fischer, A Keller, V Frinken, H Bunke, Lexicon-free handwritten word spotting using character hmms, Pattern Recognit, Lett.33(7), 934942 (2012).
14. V Frinken, A Fischer, R Manmatha, H Bunke, A novel word spotting method based on recurrent neural networks, IEEE Trans. Pattern Anal. Mach. Intell.34(2), 211224 (2012).

15. A Bensefia, T Paquet, L Heutte, A writer identification and verification system, Pattern Recognit. Lett.26(13), 20802092 (2005).
16. M Bulacu, L Schomaker, Text-independent writer identification and verification using textural and allographic features. IEEE Trans. Pattern Anal. Mach. Intell.29(4), 701717 (2007).
17. I Siddiqi, N Vincent, Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features, Pattern Recognit.43(11), 38533865 (2010).
18. ZA Daniels, HS Baird, in Proceedings of the 12th International Conference on Document Analysis and Recognition, Discriminating features for writer identification, pp.13851389, (2013).
19. R Jain, D Doermann, in Proceedings of International Conference on Document Analysis and Recognition. Offline writer identification using k-adjacent segments, pp. 769773, (2011).
20. RP dos Santos, GS Clemente, TI Ren and GDC Cavalcanti, Text line segmentation based on morphology and histogram projection, in Proceeding of the 10th International Conference on Document Analysis and Recognition, pp. 651655, (2009).
21. M Zimmermann and H Bunke,Automatic segmentation of the iam off-line database for handwritten english text,in Proceedings of the 16th International Conference on Pattern Recognition,pp. 3539, (2002).
22. D Salvi, J Zhou, J Waggoner and S Wang, Handwritten text segmentation using average longest path algorithm,in Proceedings of IEEE Workshop on Applications of Computer Vision, pp. 505512 ,(2013).
23. S Gunter, H Bunke, Ensembles of classifiers for handwritten word recognition. Int. J. Doc. Anal. Recognit.5(4), 224232 (2003).
24. H Bunke, S Bengio and A Vinciarelli, Offline recognition of unconstrained handwritten texts using hmms and statistical language models. IEEE Trans. Pattern Anal. Mach. Intell.26(6), 709720 (2004).
25. P Dreuw, P Doetsch, C Plahl and H Ney, in Proceedings of the 18th IEEE International Conference on Image Processing. Hierarchical hybrid mlp/hmm or rather mlp features for a discriminatively trained gaussian hmm: A comparison for offline handwriting recognition, pp. 35413544, (2011).
26. B Gatos, I Pratikakis and SJ Perantonis, in Proceedings of 18th International Conference on Pattern Recognition, 2. Hybrid off-line cursive handwriting word recognition, pp. 9981002, (2006).
27. E Augustin, J Brodin, M Carre, E Geoffrois, E Grosicki and F Preteux, in Proceedings of International Workshop on Frontiers in Handwriting Recognition. Rimes evaluation campaign for handwritten mail processing, pp. 231235, (2006).
28. F Montreuil, E Grosicki, L Heutte and S Nicolas, in Proceedings of the 10th International Conference on Document Analysis and Recognition. Unconstrained handwritten document layout extraction using 2d conditional random fields, pp. 853857, (2009).
29. C Kermorvant and J Louradour, in Proceedings of International Conference on Frontiers in Handwriting Recognition. Handwritten mail classification experiments with the rimes database, pp. 241246, (2010).
30. L Guichard, AH Toselli and B Couasnon, in Proceedings of International Conference on Frontiers in Handwriting Recognition. Handwritten word verification by svm-based hypotheses re-scoring and multiple thresholds rejection, pp. 5762, (2010).
31. R Wilkinson, J Geist, S Janet, P Grother, C Burges, R Creecy, B Hammond, J Hull, N Larsen, T Vogl and C Wilson, The First Census Optical Character Recognition Systems Conference (The U.S. Bureau of Census and the National Institute of Standards and Technology, 1992).
32. TM Ha and H Bunke, Off-line, handwritten numeral recognition by perturbation method. IEEE Trans. Pattern Anal. Mach. Intell.19(5), 535539 (1997).
33. SJ Smith, MO Bourgoin, K Sims and HL Voorhees, Handwritten character classification using nearest neighbor in large databases. IEEE Trans. Pattern Anal. Mach. Intell.16(9), 915919 (1994).
34. F Kleber, S Fiel, M Diem and R Sablatnig, in Proceedings of the 12th International Conference on Document Analysis and Recognition. Cvl-database: An off-line database for writer retrieval, writer identification and word spotting, pp. 560564, (2013).
35. S Fiel and R Sablatnig, in Proceedings of the 12th International Conference on Document Analysis and Recognition. Writer identification and writer retrieval using the fisher vector on visual vocabularies, pp. 545549, (2013).
36. M Diem, S Fiel, A Garz, M Keglevic, F Kleber and R Sablatnig, in Proceedings of 12th International Conference on Document Analysis and Recognition. Icdar 2013 competition on handwritten digit recognition (hdrc 2013), pp. 14221427, (2013).
37. S Al-Maadeed, D Elliman, CA Higgins, A data base for arabic handwritten text recognition research. Int. Arab J. Inf. Technol.1:, 117121 (2004).
38. S Al-Maadeed, D Elliman, CA Higgins, in Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition. A data base for arabic handwritten text recognition research, pp. 485489, (2002).
39. S Al-Madeed, C Higgins, D Elliman, Off-line recognition of handwritten arabic words using multiple hidden markov models. Knowledge-Based Syst.17(2), 7579 (2004).
40. S Al-Maadeed, Text-dependent writer identification for arabic handwriting. J. Electr. Comput. Eng (2012).