A COMPREHENSIVE HANDWRITTEN IMAGE CORPUS OF ISOLATED PERSIAN/ARABIC CHARACTERS FOR OCR DEVELOPMENT AND EVALUATION

Sara Khosravi, Farbod Razzazi, Hamideh Rezaei, Mohammad Reza Sadigh
Payasoft Company, Tehran, Iran
{Sara_khosravi, Razzazi, Sadigh} @payasoft.com, HamidehRezaee@gmail.com

ABSTRACT

In this paper, specifications, design and implementation issues of a comprehensive corpus of capital isolated handwritten character images for Persian/Arabic languages are reported. The corpus has been designed for both OCR development and evaluation purposes. The corpus contains more than 10 million characters with appropriate image quality and is supported with rich standard ground truth formatted metadata. Evaluating the accuracy of the corpus has revealed that more that 99.9% of the images are correctly labeled and the quality of more than 99.5% of images are suitable for OCR development and evaluation. This corpus may be used as a standard benchmark for OCR in Persian/Arabic OCR system.

1. INTRODUCTION

A standard formatted public OCR image corpus is an essential piece in automatic data entry and digital library systems development. They help developers train reliable and robust OCR systems and are the main tools for analysis of errors and revealing the strength and weaknesses of a system. In addition, a suitable corpus makes it possible to compare different OCR products. Therefore, a rich and publicly accepted corpus would be a good reference and benchmark in evaluation and ranking the developed OCR systems.

There is a wide variety of published OCR image corpora in the literature. [1-5]. Although there are well known Latin corpora for OCR development and evaluation[4,5], the need of preparing OCR corpora for non-Latin writings has been emphasized due to epidemic growth of computer based automation systems in the world [6]. Hence, in recent years, OCR Corpora have been publicly published in many non-English / non-Latin languages [1][7-10].

Although there are some works in Persian/Arabic typed texts Images [9] and some limited studies on cursive Arabic handwriting corpus development [11], no comprehensive reference corpus has been presented in Persian/Arabic handwritings [12]. In this paper, Hadaf corpus, as a rich and comprehensive corpus of scanned images of Persian isolated handwritten characters, is presented. The motivation of preparing this corpus is to make a reference benchmark as both a developing tool and an evaluating and ranking prerequisite for Persian/Arabic OCR industry and researchers. This corpus covers most of the data entry applications in Middle East and North African countries. The image database is well accompanied with a rich ground truth formatted meta-information, which describes the source, text, author, and other useful information of each image.

Following this introduction, in section 2, the components of Hadaf corpus is presented. Section 3 is dedicated to design and implementation procedure of the corpus. The technical specifications and the accuracy evaluation results of the corpus are presented in section 4 and 5 respectively. The paper is concluded in section 6.

2. THE CORPUS COMPONENTS

Hadaf corpus consists of three major parts. The hearth of the corpus is the image database which includes more than 10 million scanned images of handwritten capital characters in 300 dpi grayscale format. The characters have been derived from about 420000 real world data entry forms. About 2 millions of these characters are handwritten digits, while the rest are handwritten capital characters. To guarantee the originality, the image files have been stored in non-lossy bitmap format with no preprocessing.

In addition to image files, each group of the characters is accompanied with a rich XML Meta information file. There is an individual record in the XML file for each character image to describe its specification.

The corpus description, the tables of symbols, the decompression software, and user's guide are other components of the corpus.

3. DESIGN AND IMPLEMENTATION METHODOLOGY

Collection and organization of such a huge volume of data with documented metadata is neither manually feasible, nor reliable due to human errors. On the other hand, there is no full automatic mechanism without OCR errors. In addition, the data entry forms include some

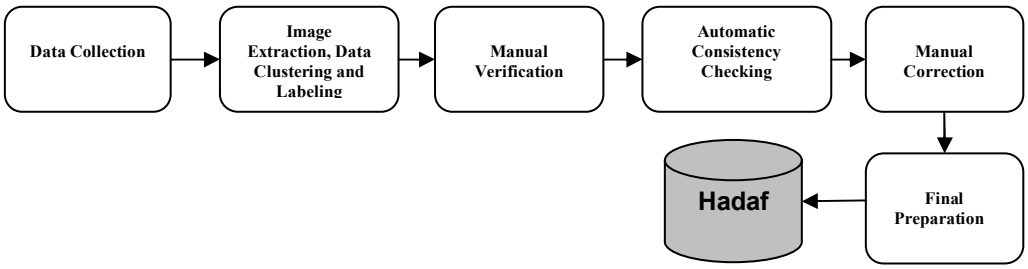

Figure 1. Hadaf Corpus preparation Block Diagram

form filling errors which make the derived database unreliable. Hence, a combined manual-automatic methodology was designed to make a trade off between the preparation efforts versus corpus accuracy.

The block diagram of the procedure is shown in figure 1. At the first stage, the data entry forms have been spread out and filled nationwide in two subsequent years. All of the forms have been gathered and scanned in a central site. The forms have been passed through a preprocessing stage including form alignment and character image extraction procedures. In addition, the OCR engine has been applied to all characters to find out the equivalent text of each extracted image. To have a certified registration procedure, all of the recognized fields of the forms have been verified manually by operators. In the consistency check stage, the modified characters of all corrected fields are separated automatically in a set which should be rechecked by operators in the next stage. This mismatch is created due to not only the machine recognition errors, but also the form filling and the operator errors. At the final stage, an XML database has been prepared and the images and their corresponding meta-data has been compressed and packed. In the following, each of the stages of corpus preparation is presented in more detail.

3.1. Data Collection

Data is collected from registration forms of NODET[1] nationwide qualification test which is conducted each year among the students of last year of primary schools. Almost 220,000 students participate in this contest. All the forms have been filled by the school officers or parents of the students. Hence a wide variety of handwritings are gathered in this registration. All fields of these forms, including personal and address information of the applicants, are filled in capital isolated Persian characters. The forms have been scanned by document scanners in the central site of the NODET and these images have been stored in 300dpi, 256 gray scaled bitmap format.

3.2. Image Extraction, Data Clustering and Labeling

In the central site, the character images have been extracted after automatic form alignment procedure. A few samples of character images are shown in figure 2. The extracted characters have been labeled and clustered by the results of automatic recognition of the forms fields by an OCR engine in this stage. Certainly, this metadata is not error free and should be verified in various aspects. The remaining blocks of corpus preparation sequence performs the mentioned metadata correction.

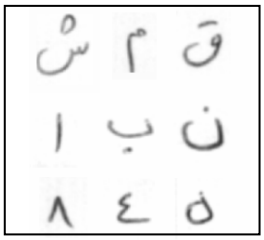

Figure 2. Hadaf Corpus character images samples

3.3. Manual Verification

All the recognized fields have been checked out to compensate OCR and form filling errors. The output of this stage is a database of corrected fields of the forms with less than 2% erroneous forms.

3.4. Automatic Consistency Checking

In the foregoing stage, some of the characters in the fields have been corrected by the operator. Although this correction is acceptable for the registration procedure, there are some errors whose source is not OCR engine (e.g. misunderstanding the form by the author). This kind of errors makes the metadata fields of the characters confused and should be corrected in the corpus. To avoid rechecking the whole data, all of the modified characters in the previous stage have been gathered in a separate set. This set is about 15% of the whole database.

[1] National Organization for Development of Exceptional Talents

```
>z:row CharName='01-F091605019001077487-State3' CharFolder='CD-DEMO\OCR1-H-CD34\01' Class='01' Gender='د' City = 'مشهد'
ExtractedField='State' CharNumber='3' Scanner='Fujitsu 3093GX' CharWriterIndex='RF091605019 ' TrainOrTest='Train</'
>z:row CharName='01-F091605019001077487-State5' CharFolder='CD-DEMO\OCR1-H-CD34\01' Class='01' Gender= 'د' City= 'مشهد'
ExtractedField='State' CharNumber='5' Scanner='Fujitsu 3093GX' CharWriterIndex='RF091605019' TrainOrTest='Train</'
```

Figure 3. Meta-information XML file Sample of Character Images

3.5. Manual Correction

The characters with inconsistent data were rechecked to be categorized in the correct character class. About 20% of this set were re-categorized.

3.6. Final Preparation

3.6.1 Final XML database preparation

The meta-information of all characters has been formatted in a unified ground truth XML format [13- 15]. A sample of corpus meta-information is presented in figure 3.

3.6.2 Data Compression and packing

There are 10236040 sample characters in the image database, which have been divided into 51 sets with approximately 200,000 characters in each set. The size of each set is about 2.4G bytes, which is compressed to be stored in a Compact disk. The compression format is the freeware bz2 compression format.

3.6.3 Documentation

At the last step, the documentation of the corpus has been prepared in the OCR standard corpus format [14] and has been added to the corpus. The documents include a user guide and manual, a specification sheet and required character-code conversion tables.

4. TECHNICAL FEATURES

The main technical specifications of Hadaf Corpus are tabulated in table 1. The specification reveals that Hadaf is the largest and comprehensive corpus for Persian/Arabic OCR development and benchmarking. Although the corpus is large enough to fairly evaluate different OCR products, the method of storing images and their metadata helps the user choose and use a subset of the corpus.

In this corpus, each image is accompanied with rich meta-information including image file information, author information, the parent word which the character image is extracted from, imaging device information and a statistically suitable train set/ test set label which helps developers to train and evaluate statistical learning machines reliably. Table 2 shows the meta-information of each image in detail. A sample of the XML formatted ground-truth meta-information is presented in figure 3.

5. CORPUS ACCURACY EVALUATION

The meta-information of the image is extractable automatically from the data entry procedure sequence with no errors. The only field of meta-information which may tolerate some errors is the image equivalent text.

Table 1. Hadaf Corpus Specifications

Specification	Value
Name	Hadaf-84
Date of Release	Oct. 2005
Character Type	Capital Isolated Persian Digits and Letters
Number of Samples	10236040
Storing Size	120GB
Source	Registration forms of NODET
Imaging Specifications	300dpi, 256 gray scaled imaging
Image file type	bitmap
Compression format	bz2
Image Size	52.6mm*48.0mm (77*95Pixels)
Train/Test Ratio	70%/30%
Storing method	Divided into 200000 sub-corpora, Separate folders for each character class
Meta-information format	XML ground truth
Characters Statistics	According to data entry registration forms

Table 2. Meta-information of each character image

Meta-Information Group	Meta-Information Title
Image File Information	File Name
	File Path
	Image Text
Author Information	Author Number
	Gender
	City
Source word Information	Field Name
	Character Number in the field
Imaging Information	Scanner Type
Train set/Test set categorization	Train Set/ Test Set label

In addition, the image text is the most important field of the character images.

Therefore, to verify the corpus categorization accuracy, an evaluation test has been performed. In this test, a set of 133529 random selected images of the corpus has been selected. To select the character images appropriately, 47 Persian/Arabic characters, including both letters and digits, are clustered into 14 classes. The handwritings of the elements of each cluster are nearly similar. Major intra-cluster differences are the number of dots or slight handwriting variations. Therefore the artifacts, noises and human categorization errors seem similar in each cluster. The number of elements of each cluster is determined according to Persian language characters statistics. In addition, the samples are selected uniformly over all corpus subsets.

To evaluate the corpus accuracy, all errors are counted by the operators in the extracted samples. The errors may be divided into two main subclasses. The first subclass points to sever errors, containing missing dots or other parts of the character, cutting the main part of the character, a sever noise or cross out which makes the character unreadable, artifacts caused by bad imaging or mis-categorization. In the second subclass, minor errors, containing minor cut character images and minor noises or cross outs in character images are gathered. The selection criterion of minor errors is selecting the error characters that the presence of them in the corpus would not affect the evaluation or training performance of the corpus.

The evaluation results show that in 133529 character images, there are 567 images (0.42%) with sever errors, containing 115 (0.08%) categorization errors. In addition, the number of minor errors are 716 (0.54%) in evaluation samples in the future.

6. CONCLUSTION

Reliable OCR image corpora have a major role in OCR applications development and its world wide deployment. Hadaf as a comprehensive handwritten character image corpus with rich and accurate meta-information is a good tool for OCR application development in Middle East and North African countries. Enriching the corpus with handwritings of several countries and different situations of writing may help the corpus to be more compatible with real world applications.

7. ACKNOWLEDGEMENT

Authors would like to acknowledge the financial support provided by Iranian Supreme Council of Information and Communication Technology. In addition, authors would like to acknowledge the permission of NODET to use their valuable registration data.

REFERENCES

[1] S. Mihov, K. Schulz, Ch. Ringlstetter, V. Dojchinova, V. Nakova, K. Kalpakchieva, O. Gerasimov, A. Getcharek, C. Gercke, "A corpus for Comparative Evaluation of OCR Software and Post correction Techniques", *in Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR 2005),* Aug 2005, Seoul, Korea, Vol. 1, PP. 162-166.

[2] S. Rice, F. Jenkins, T. Nartker, "The fifth annual test of OCR accuracy", *Technical Report TR96-01,* Information Science Research Institute, University of Nevada, Las Vegas, 1996.

[3] T. Philips, S. Chen, R. Haralick, "A database of document images for OCR Research", *University of Washington, English/ Japanese document image database,* CDROM, 1995.

[4] http://www.nist.gov/srd/optical.htm

[5] http://hwr.nici.kun.nl/unipen

[6] I. Guyon, R. Haralick, J. Hull, I. Philips, "Data sets for OCR and document image understanding research" *in Handbook of Character Recognition and Document Image Analysis,* World Scientific, 1997, PP. 779-799.

[7] T. Kanungo, M. Kamiya, O. Bulbul, "Project Guttenberg, Bible Image Corpus for Multilingual OCR Evaluation and Training", *Technical Report,* University of Maryland, College Park, 1998.

[8] E. Kavallieratou, N. Liolios, E. Koutsogeorgos, N. Fakotakis, G. Kokkinakis, "A Greek Database of Unconstrained Handwriting", *in Proceedings of IEEE Conference on Pattern Recognition,* Sep. 2000, Barcelona, Spain, PP.

[9] R. Davidson, R. Hopeley, "Arabic and Persian OCR training and test data sets", *in Proceedings of Symposium of Document Image Understanding Technology (SDIUT 97),* Annapolis, May. 1997, PP. 303-307.

[10] J. Baker, A. Hardlie, A. McEnery, H. Cunningham, R. Gaizauskas, "EMILIE, A 6-7 Million Word Corpus of Indic Languages: Data Collection, Mark up and Harmonization", *in Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC 2002),* Las Palmas, May 2002.

[11] Y. Ohali, M. Cheriet, C. Suen, "Databases for recognition of handwritten Arabic cheques", *in Pattern Recognition,* Vol. 36, PP. 111-121.

[12] L. Lorigo, V. Govindaraju, "Offline Arabic Handwritten Recognition, A Survey", *in IEEE Transactions of Pattern Analysis and Machine Analysis,* Vol. 28-5, PP. 712-724.

[13] M. Agrawal, K. Bali, S. Madhvanath, "UPX: A New XML Representation for Annotated Datasets of Online Handwriting Data", *in Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR 2005),* Aug 2005, Seoul, Korea, Vol. 2, 1161- 1165.

[14] A. Bhaskarabhata, S. Madhvanath, "An XML Representation for Annotated Handwriting Databases for online Handwriting Recognition", *in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004),* Lisbon, Portugal, May 2004.

[15] T. Kanungo, C.H. Chen, J. Czorapinski, I. Bella, "TRUEVIZ, A Ground-truth/metadata editing and visualization toolkit for OCR", *in proceedings of SPIE Conference on Document Recognition and Retrieval,* Jan. 2001.