IfN/Farsi-Database: A Database of Farsi Handwritten City Names

Saeed Mozaffari

Faculty of Engineering,
Semnan University,
Semnan, Iran
e-mail:
mozaffari@semnan.ac.ir

Haikal El Abed,
Volker Märgner

Institute for
Communications
Technology (IfN),
Braunschweig Technical
University, Germany
e-mail:
{elabed,v.maergner}@tu-bs.de

Karim Faez,
Ali Amirshahi

Electrical Engineering
Department
Amirkabir University of
Technology, Tehran, Iran
e-mail:
{kfaez,s.a.amirshahi}@aut.ac.ir

Abstract

The great success and high recognition rates of both OCR systems and recognition systems for handwritten words are inconceivable without the availability of huge data sets of real world data. For Arabic handwritten recognition systems, the availability of the IFN/ENIT-database provides more possibilities to compare the performance of several systems, motivated by this fact we have developed a new database with the same characteristics of the IFN/ENIT-database. This make the adaptation of existing system for the Farsi handwritten words recognition easier. In this paper we present IfN/Farsi-database, a new database of handwritten Farsi words. The database consists 7,271 binary images of 1,080 Iranian province/city names, collected from 600 writers. For each image in the database, the ground truth information including its ZIP code, sequence of characters, and numbers of characters, sub-words and dots are also available. This database is available for research and academic use.

Keywords: Arabic/Farsi OCR, database, IFN/ENIT-database, IfN/Farsi-database.

1. Introduction

Latin OCR seems to reach maturity. For many practical applications such as postal automation and bank cheque reading systems current OCR systems are capable of recognizing handwritten forms, words and even sentences, very fast and accurate [12]. This advancement is due in part to several large data sets such as CEDAR [9] , NIST [26], CENPARMI [25], UNIPEN [8], etc.

The first step in any OCR research project is collecting a data set with enough samples which can properly model real application environment. Standard databases play crucial rule in development, evaluation and comparison of different recognition systems. Free available comprehensive and standard databases obviate the need to collect data from each research group separately. As a result, they can allocate more time for algorithm development rather than data collection. By reviewing the literature, it is obvious that most of Arabic/Farsi OCR researchers implemented their system on set of data gathered by themselves and different recognition rates were reported [5, 4, 3, 19]. Therefore, it is very difficult to objectively compare the results for the proposed methods.

Large datasets serving as baseline databases to test recognition systems is crucial for performance evaluation [15]. Another crucial point is the database specialized features to solve very diverse tasks, as the recognition rate, the common measure to assess system performance, is not really relevant for system component development [7, 6]. To improve the overall system quality, it is essential to know the effectiveness of their features. In filed of Arabic handwriting recognition, progresses or reviews of state-of-the-art algorithms are reported in many papers, e.g.,[13], or even in several special sessions or workshops. A major reason for the advances in this field is the availability of the IFN/ENIT-database [21] as a de facto standard for training, testing, and comparing different systems and different algorithms for Arabic handwritten word recognition. The IFN/ENIT-database was intentionally developed to advance the research and development of Arabic handwritten word recognition systems. Since the presentation of this database at the CIFED 2002 conference [21], until beginning of 2007), more than 62 groups in about 31 countries are working with the IFN/ENIT-database, which is freely available (www.ifnenit.com) for noncommercial research.

Recently, for Arabic OCR, several alternative

databases are published, such as for handwritten texts [1], machine-printed documents [23], handwritten words [21] and bank checks [2]. In addition to these data sets, some free databases were also presented for Farsi digits and characters [17, 24, 11]. To the knowledge of the authors, currently no public database containing handwritten Farsi words is however available.

In this paper we present IfN/Farsi-database, which consists of 7,271 handwritten Farsi words produced by approximately 600 writers. The underlying lexicon includes 1,080 different words selected from the list of Iranian province/city names. The truth file for each lexicon entity can be used to train and test word recognizers. Furthermore, it particularly contains useful information for lexicon reduction techniques. The database described in this paper will be made available for academic use upon request. In the next section, we describe the database acquisition procedure. Section 4 describes the ground truth of the database. Section 5 describes shortly the verification and post-processing steps for the data. In Section 6 and 7 further statistics of the collected data are listed. Finally some conclusion remarks are presented in Section 8.

2. Overview of the IfN/Farsi-database

Due to some ambiguities and large diversities of writing styles, recognition systems are generally based on a set of possible words called lexicon. Depending on the application type, the size of the lexicon can vary from 20-30 words (e.g., in reading of check amounts [10]) to 10,000-60,000 words (e.g., for English text recognition [12]). In case the lexicon consists of more than 1000 classes (words), it is usually considered as a large lexicon [12]. By increasing the lexicon size, recognition accuracy and recognition time decreases and increases respectively.

The IfN/Farsi-database has a large lexicon with 1,080 words selected from the list of Iranian province/city names. Approximately, 600 persons with different ages and different educational backgrounds participated in this data collection. Each writer was asked to fill at most two forms comprising 24 pre-selected province/city names and their corresponding postcode. However, the postcodes have not been processed yet. An example of a filled form is shown in Figure 1. After collecting all forms, the province/city name and the respective postcode fields were extracted automatically. Based on the label string at the bottom of each form, which shows the list of pre-selected province/city names, the ground truth information were created automatically. This information includes each word's characters sequence, the number and the position of its dots, and the number of sub-words (PAWs). The whole IfN/Farsi-database consists of 7,271 handwritten Farsi word images.

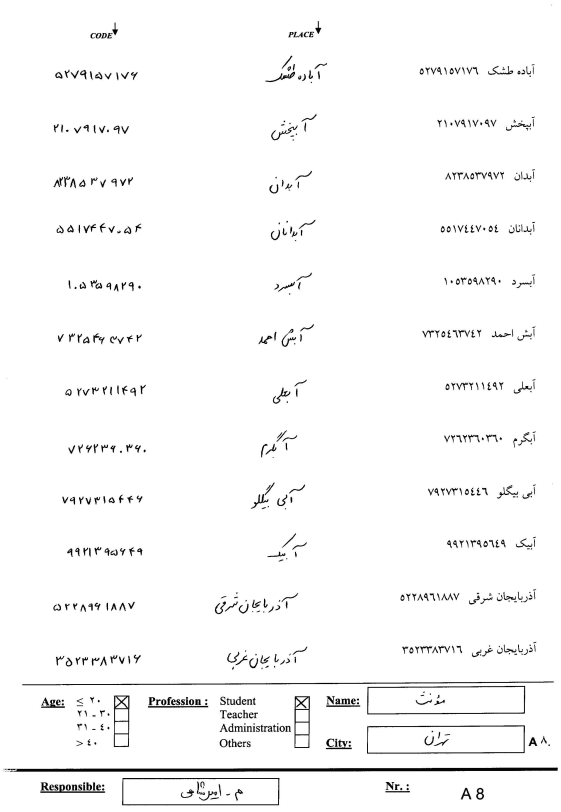


Figure 1. An example of a field form.

3. Form Design and Data Collection

The form was designed in such a way to:

- collect data without strong constraints
- collect handwritten city names written in a similar quality as those on the address field of a letter
- be easy to be processed automatically
- provide additional information about the person, who filled it.

As it is shown in Figure 1, each form consists of three columns on the top and a text box on the bottom. On the top section of each form, 12 Iranian city names and a randomly generated digit string were printed in separate lines. The city names were automatically chosen in the forms in such a way that all city names have the same number of appearance in the whole dataset. Digits also have a uniform distribution. However, in IfN/Farsi-database, these digits have not included yet. The writer's age, profession and gender were added to the text box on the bottom of the form. This information will help us to present good statistics about IfN/Farsi-database.

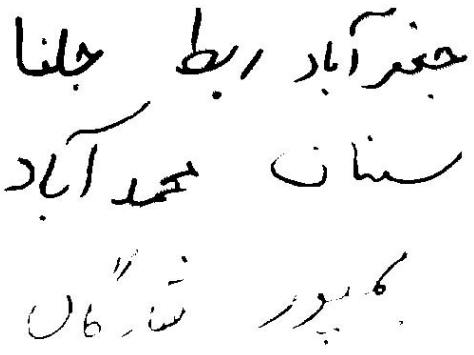
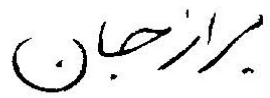


Figure 2. Effect of different writing instruments.

For data extraction convenience, each entry field on the form is commonly represented as a box. These boxes can impose some restrictions on the writer and consequently alter his/her writing style. To have an unconstrained database, we did not print any bounding box or guiding lines. As a writing guidance, we printed black rectangles on the backside of each form, which can be seen through from the front side and thus mark the writing area. In the scanning process these black rectangles can be removed using a simple threshold. Further segmentation operations are then not necessary. To get the most natural and unconstrained way of writing, the writers were asked to use their every day writing. In order to avoid any pressed and deformed words, the black rectangles of each entry were designed for the longest word in the lexicon. As writing too many words is a tedious task and can change every day writing style, only 12 words were written in each form, and we asked each writer to fill at most two forms. A page number is employed as the form identifier for the subsequent labeling process. since we did make any constraint to the writing instruments, we obtained all kinds of writing instruments in our database (Figure 2). Nevertheless, due to many broken characters we had to exclude some words from the database during the post-processing stage.

The filled forms were then scanned with resolution of 300 dpi and converted to black and white (binary) images. Because the paper was white and the words were written with a black or dark blue pen, the binarization was not a problem. During the page scanning the page number and the additional information were keyed in manually. Page slope correction was performed automatically using the black line on the bottom of the page as a horizontal reference. Finally, an advanced projection method was performed to extract the word and the postcode images on the page automatically.



(a)

```
COM: ia013_004.tif
X_Y: 285 71
BDR: begin data record
LBL: ZIP: 1149; AW1: برازجان; AW2: baB|raE|aaA|zaA|jaB|aaE|naA|
CHA: 7
PAW: 5
DOT: 4; UD: 2; DD: 2
DOT PATTERN: 1D1U1D1U
EDR: end of data record
```

(b)

Figure 3. (a) Original word images, (b) Ground truth of the word 'برَازجَان'.

4. Labeling and Ground Truth

In order to use the word images for training and testing, we need to assign labels consistently to them. Labeling process is usually expensive, time consuming and error prone. However, thanks to the form identifier each word can be automatically assigned to its corresponding label. The label of each word consists of a four-digit postal code, a word in Arabic with code set ISO 8859-6, and a code that describes the sequence of the character shapes. This character shape code is generated automatically and provides character shape occurrence information. Each character is represented by a Latin string, separated from each other by a vertical line "|". In Farsi and Arabic handwriting the shape of each character varies according to its position in the word. To make the code more readable, we added an additional Latin character as an index to show the character position. 'B' stand for beginning, 'M' for middle, 'E' for end, 'A' for alone/isolated character shapes, and 'L' for the ligature. The label of the word begins with 'LBL' identifier. Figure 3 shows an example of the label and the ground truth of a word in the IfN/Farsi-database.

In addition to the word's label, the following information was added to the ground truth file:

- name of the corresponding image file (COM).
- image size (X_Y).
- the number of characters (CHA).
- the number of PAWs (PAW).
- the number of dots (DOT), up dots (UD), and down dots (DD).
- dots pattern (DOT PATTERN).

For each dot(s), a two-character string is assigned as nU or nD, in which character n shows its groups (one,

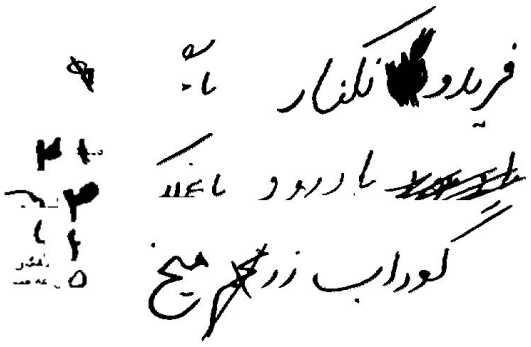


Figure 4. Some examples of discarded samples.

two or three) and characters U and D indicate *up* and *down* positions respectively. Subsequently, these labels are concatenated to each other to produce the holistic dots pattern. For example, the dots pattern corresponding to the word 'شیراز' is '3U2D1U'. This dots pattern can be used as a shape descriptor for other applications like lexicon reduction [18, 20].

5. Verification and Postprocessing

The filled forms usually have different types of error. For example, they may contain altered, displaced or empty boxes. Variability of handwriting, different writing instruments or simply writing errors are other sources of error. Therefore, manual verification is necessary after data extraction and labeling stages.

In the post-processing stage, all collected word images were verified by a human observer and any unreadable words or any very low quality samples were discarded from the database. Additionally the assigned label to each word was checked, because due to different writing styles adjacent characters can be combined to gather and form a new shape called ligature, which is not resemble to the original characters. Figure 4 shows some examples of discarded words.

6. Database details and statistics

This section describes the IfN/Farsi-database in more details regarding the useful statistics about the number of images, words, PAWs, characters, and writers.

Age, experience, and education are important factors which affect the quality of handwritten words. For this database we tried to collect data from peoples with different ages and different educational backgrounds. Approximately 60% of the writers were men and the rest were women. Table 1 shows the distribution of writers of the IfN/Farsi-database.

The IfN/Farsi-database consists of 7,271 samples from 1,080 Iranian city names. Figure 5 shows the distribution of 1,080 words based on the number of characters. According to figure 5, the longest word in the IfN/Farsi-database has 17 characters. The PAWs and the using of many dotted characters make Arabic/Farsi languages unique. Besides, the number of characters, the PAWs number, and the dots number and the position information can be used for some applications, e.g. Arabic/Farsi lexicon reduction [18] and word spotting. The 1,080 classes in IfN/Farsi-database can also be categorized according to the number of sub-words (PAWs), the dots number, and the dots position. Figures 6, 7 and 8 show these distributions. The dots number and position of the database entities are represented likely as uniform distribution and consequently the degree of lexicon reduction will be increased [19]. Figure 9 shows those classes that contains more than 10 city names. The IfN/Farsi-database also includes 23,545 subwords and 43,501 characters.

Table 1. Writers age distribution.

Age	Percent
less than 20	30
between 20 and 30	50
between 30 and 40	10
more than 50	10

By increasing the lexicon size, the ambiguity increases due to the presence of more similar words in the lexicon and it leads to more confusion to the classifiers. Because of the frequent use of dots in Arabic and Farsi languages, some of lexicon entries (city names) look very similar to each other, which can be distinguished only by considering their dots number or position. Figure 10 shows some of these similar words.

7. Working with the IfN/Farsi-database

The IfN/Farsi database of handwritten Iranian province/city names is freely available for research and academic use. It can also be ordered by sending an e-mail to the corresponding author. At the moment, we are adapting the IfN handwritten Arabic word recognition system [22, 16, 14] to work with the IfN/Farsi-database. The recognition results will be published soon.

Since the IfN/Farsi-database has a large lexicon and also a comprehensive information in its ground truth, similar lexicon reduction techniques used for Arabic language [18, 20] can be applied to this database. As IfN/Farsi-database is unconstraint regarding writing instruments, many preprocessing approaches can be applied depending on the back-end application.

Taking the fact that Farsi and Arabic languages are very similar to each other, we hope that presenting such databases can bring many international benefits.

8. Summary and Conclusion

In this paper we present the IfN/Farsi-database as a new database for handwritten Arabic words. The IfN/Farsi-database is publicly available for the purpose of research. With 600 different writers a wide variety of writing styles is guaranteed. After verification and post-processing steps, 7271 word images were selected from the list of 1,080 Iranian province/city names. For each word image in the database a corresponding ground truth is available. Ground truth information can be useful for training and testing of any handwriting word recognizer. Furthermore, ground truth information was structured in such a way that it can be used for lexicon reduction algorithms.

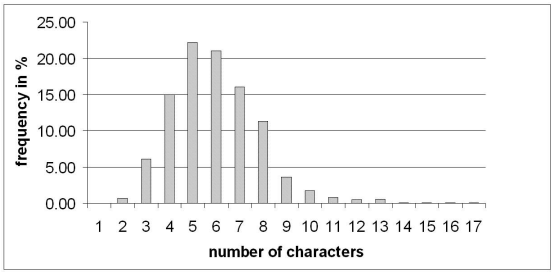


Figure 5. Distribution of the IfN/Farsi words based on the number of characters.

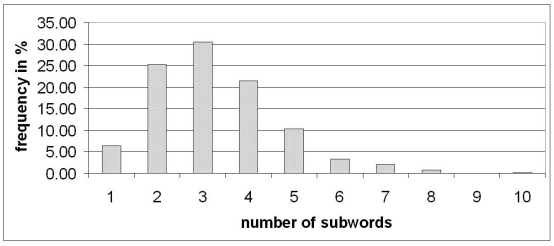


Figure 6. Distribution of the IfN/Farsi words based on the number of subwords.

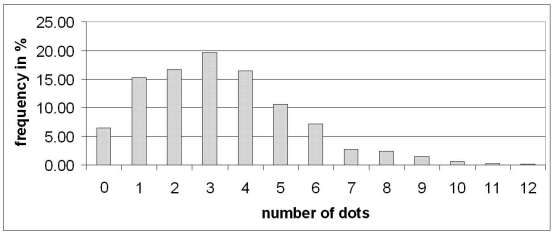


Figure 7. Distribution of the IfN/Farsi words based on the number of dots.

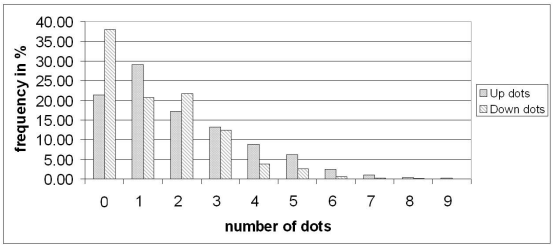


Figure 8. Distribution of the IfN/Farsi words based on the position of dots.

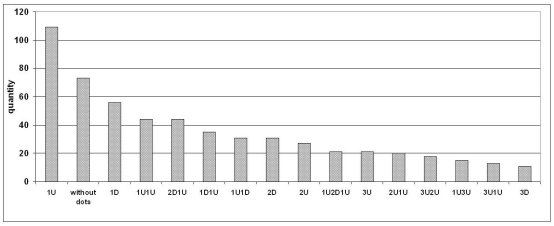


Figure 9. Dense classes in the IfN/Farsi with more than ten words.

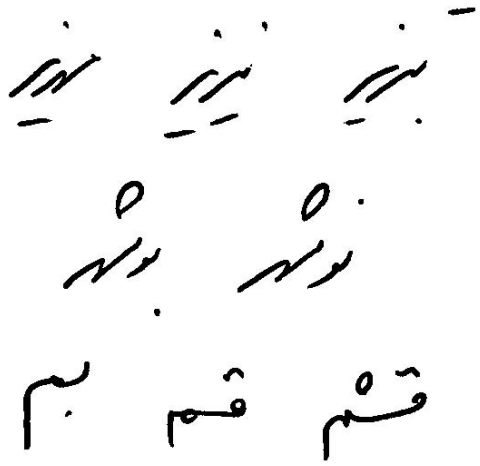


Figure 10. Examples of similar words in the IfN/Farsi-database.

References

[1] S. Al-Ma'adeed, D. Elliman and C. Higgins, "A data base for Arabic handwritten text recognition research", D. Elliman, editor, *8th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 2002, pp 485–489.

[2] Y. Al Ohali, M. Cheriet and C. Y. Suen, "Databases for recognition of handwritten Arabic cheques.", *Pattern Recognition*, 36(1):111–121, 2003.

[3] A. Broumandnia and J. Shanbehzadeh, "'Fast Zernike wavelet moments for Farsi character recognition"', *Image and Vision Computing*, 25(5):717–726, 2007.

[4] M. Dehghan, K. Faez, M. Ahmadi and M. Shridhar, "Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM ", *Pattern Recognition*, 34(5):1057–1065, 2001.

[5] M. Dehghan, K. Faez, M. Ahmadi and M. Shridhar, "Unconstrained Farsi handwritten word recognition using fuzzy vector quantization and hidden Markov models", *Pattern Recognition Letters*, 22(2):209–214, 2001.

[6] H. El Abed and V. Märgner, "Comparison of Different Pre-Processing Methods for Offline Recognition of Handwritten Arabic Words", 9^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, 2007, volume 2, pp 974 –978.

[7] H. El Abed and V. Märgner, "The *IFN/ENIT*-Database - a Tool to Develop Arabic Handwriting Recognition Systems", *IEEE International Symposium on Signal Processing and its Applications (ISSPA)*, 2007.

[8] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman and S. Janet, "UNIPEN project of on-line data exchange and recognizer benchmarks", *12th International Conference on Pattern Recognition (ICPR)*, 1994, volume 2, pp 29–33.

[9] J. Hull, "A database for handwritten text recognition research", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

[10] G. Kaufmann, H. Bunke and T. M. Ha, "Recognition of cursively handwritten words using a combined normalization/perturbation approach", *5th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 1996, pp 17–22.

[11] H. Khosravi and E. Kabir, "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties", *Pattern Recognition Letters*, 28(10):1133–1141, 2007.

[12] A. L. Koerich, R. Sabourin and C. Y. Suen, "Large vocabulary off-line handwriting recognition: A survey", *Pattern Analysis & Applications*, 6(2):97–121, 2003.

[13] L. Lorigo and V. Govindaraju, "Offline Arabic handwriting recognition: a survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):712–724, 2006.

[14] V. Märgner and H. El Abed, "ICDAR 2007 Arabic Handwriting Recognition Competition", *9th International Conference on Document Analysis and Recognition (ICDAR)*, 2007, volume 2, pp 1274–1278.

[15] V. Märgner and H. El Abed, *Arabic and Chinese Handwriting Recognition,* , volume 4768, chapter Databases and Competitions: Strategies to Improve Arabic Recognition Systems, pp 82–103, Springer, LNCS, 2008.

[16] V. Märgner, M. Pechwitz and H. El Abed, "ICDAR 2005 Arabic Handwriting Recognition Competition", *8th International Conference on Document Analysis and Recognition (ICDAR)*, 2005, volume 1, pp 70–74.

[17] S. Mozaffari, K. Faez, F. Faradji, M. Ziaratban and S. M. Golzan, "A Comprehensive Isolated Farsi/Arabic Character Database for Handwritten OCR Research", *10th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 2006, pp 385–389.

[18] S. Mozaffari, K. Faez, V. Margner and H. El-Abed, "Strategies for Large Handwritten Farsi/Arabic Lexicon Reduction", K. Faez, editor, *9th International Conference on Document Analysis and Recognition (ICDAR)*, 2007, volume 1, pp 98–102.

[19] S. Mozaffari, K. Faez, V. Märgner and H. El Abed, "Lexicon Reduction Using Dots for Off-line Farsi/Arabic Handwritten Word Recognition", *Pattern Recognition Letters*, In press, 2008.

[20] S. Mozaffari, K. Faez, V. Märgner and H. El Abed, "Two-Stage Lexicon Reduction For Off-Line Arabic Handwritten Word Recognition", *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, In press, 2008.

[21] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze and H. Amiri, "IFN/ENIT-Database of Handwritten Arabic Words", *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, 2002, pp 127–136.

[22] M. Pechwitz and V. Märgner, "HMM based approach for handwritten arabic word recognition using the IFN/ENIT-database", *7th International Conference on Document Analysis and Recognition (ICDAR)*, 2003, pp 890–894.

[23] S. G. Sclosser"ERIM Arabic document database", http://documents.cfar.umd.edu/resources/database/ERIM.

[24] F. Solimanpour, J. Sadri and C. Suen, "Standard databases for recognition of handwritten digits, numerical string, legal amounts, letters and dates in Farsi language", *10th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 2006, pp 3–7.

[25] C. Suen, C. Nadal, R. Legault, T. Mai and L. Lam, "Computer recognition of unconstrained handwritten numerals", *Proceedings of the IEEE*, 80(7):1162–1180, 1992.

[26] R. A. Wilkinson, J. Geist, S. Janet, P. J. Grother, C. J. C. Burges, R. Creecy, B. Hammond, J. J. Hull, N. J. Larsen, T. P. Vogl and C. L. Wilson, "The First Census Optical Character Recognition System Conference", Technical Report NISTIR 4912, National Institute of Standards and Technology, 1992.