A New Large Urdu Database for Off-Line Handwriting Recognition

Malik Waqas Sagheer, Chun Lei He, Nicola Nobile, and Ching Y. Suen

CENPARMI (Centre for Pattern Recognition and Machine Intelligence)
Computer Science and Software Engineering Department, Concordia University
Montreal, Quebec, Canada
{m_saghee,cl_he,nicola,suen}@cenparmi.concordia.ca

Abstract. A new large Urdu handwriting database, which includes isolated digits, numeral strings with/without decimal points, five special symbols, 44 isolated characters, 57 Urdu words (mostly financial related), and Urdu dates in different patterns, was designed at Centre for Pattern Recognition and Machine Intelligence (CENPARMI). It is the first database for Urdu off-line handwriting recognition. It involves a large number of Urdu native speakers from different regions of the world. Moreover, the database has different formats – true color, gray level and binary. Experiments on Urdu digits recognition has been conducted with an accuracy of 98.61%. Methodologies in image pre-processing, gradient feature extraction and classification using SVM have been described, and a detailed error analysis is presented on the recognition results.

Keywords: Urdu OCR, Off-line Handwriting Recognition, Handwriting Segmentation, Urdu Digit Recognition.

1 Introduction

Off-line handwriting recognition has become an important area in the pattern recognition field and finding a good database for recognition is also a main issue. The writers of this Urdu database have been divided into three categories based on their genders, and writing orientations. For the Urdu language, this is the first known handwritten database and it can be used for multiple applications. The value of this Urdu database is based on the variety of its contents. The databases contain a large number of Urdu isolated digits and numeral strings of various lengths, including those with decimal points. This database can be used for various types of applications such as digit, letter, word, and cheque recognition as well as for word spotting applications.

Urdu is an Indo-European [1] language which originated in India. It is a popular language of the subcontinent. Urdu is one of the 23 official languages of India and is one of the two official languages of Pakistan. This language is also widely spoken in Dubai. It is one of the most spoken languages in the world.

Written Urdu has been derived from the Persian alphabet, which itself has been derived from the Arabic alphabet. Like Arabic, Urdu is written from right to left. However, Urdu has more isolated letters (37) than Arabic (28) and Persian (32). This fact

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 538–546, 2009.

makes Urdu different from Arabic and Persian in appearance in such a way that it uses slightly more complicated and complex script. Therefore, constructing an Urdu database for handwriting recognition was necessary.

The structure of this paper is divided into seven sections. In Section 2, we describe the data collection process. Section 3 comprises of data extraction and a description of preprocessing methods. In section 4, we discuss the overview of the database followed by descriptions of the datasets. We describe the ground truth data in Section 5. In section 6, experiments on Urdu isolated digits and their error analysis have been described. Finally, we discuss the conclusions as well as future work.

2 Data Collection Process

A two-page data entry form had been designed and used for the collection of handwritten data. The process of data collection was conducted in Montreal, Canada (30%) and Pakistan (70%). The first page of the form contains 20 Indian isolated digits (two samples of each digit), a freestyle written date, 38 numeral strings of various lengths, 43 isolated characters and 16 words. The second page includes the remaining 41 words and five special symbols. Fig. 1 shows a small portion of our form.

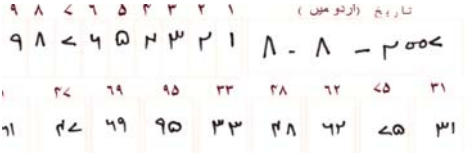


Fig. 1. Sample of Filled Form

So far, we collected handwriting samples from 343 different writers comprised of men and women from various professional backgrounds, qualifications, and ages. We were interested in keeping track of writer's gender, age, as well as whether they were right-handed or left-handed. Even though at this stage the writer information has no significance, it could be used in future research. The writers have been distributed into three categories as follows: 1) right-handed males (75.4%), 2) left-handed males (5.6%), and 3) right-handed females (19.0%). We can see that the number of left-handed subjects are far less than the number of right-handed ones.

3 Data Extraction and Preprocessing

We obtained digital copies of the forms by scanning and saving them as true color (24 Bit), 300 DPI lossless TIFF images. Basic noise removal was done at the form level before we applied our algorithm to remove the red lines from the forms. Removing these lines facilitates the extraction of the handwritten elements. Removing the lines requires careful inspection of the elements on the scanned page since some writers exceeded the field boundaries and overlapped the red line as seen in Fig. 2a-c. Furthermore, although the writers used either blue or black ink, if we zoom in some

handwriting, as seen in Fig. 2d, we can see that some pen strokes contain traces of red (in different shades of red). Therefore, our algorithm had to carefully remove red lines without affecting the writing. Doing so would introduce salt and pepper noise and lose the true outline of the handwritten element.

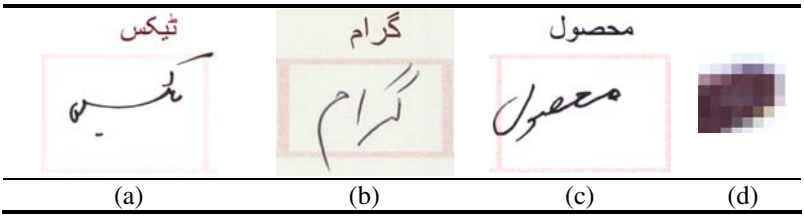


Fig. 2. Examples of Urdu Handwritten Elements Exceeding the Field Boundaries (a-c), and Close-up of Handwriting Displaying Several Colors

Removing the red lines involves analyzing each relevant pixel in the form. For each of these pixels, we determine if it should be removed by determining if its "redness" is within a range of red we had used for printing the forms. We then look at the neighboring pixels – if they are also red within the same range, then it could be potentially removed. If at least one bordering pixel is of handwriting colour, normally black or blue, then the pixel is kept, otherwise, it is considered a border pixel and is removed (replaced by white).

Several forms contained unwanted artifacts because of folding and crumpling of the paper mostly due to the handling by postal delivery from Pakistan to Canada. In addition, a number of forms had an overall dark background. In both cases, these problems had to be corrected before the forms were converted to grayscale. The cleaned forms were then passed to the handwriting extraction process.

For each handwritten sample, the box's coordinates were located. After extracting all the handwritten samples, a special filter was applied to remove the salt and pepper noise. Also, there was a great effort to manually clean those images. All the form images have been saved in three formats: true color, gray scale and binary. A special program was developed to automate the data extracting process from the true color forms. This program takes the coordinates of each box and extracts the box image in true color. The image from the same box numbers of each form is saved in a unique folder. All extracted handwritten items were then converted to grayscale and binary.

4 Database Overview

After creating the databases in true color, it was converted into gray scale and binary formats. Each database had six basic datasets which consisted of isolated digits, dates, isolated letters, numeral strings, words, and special symbols. For each dataset, the data was divided into training, validation, and testing sets by picking elements in the data using a ratio of 3-1-1. This resulted in an approximate distribution of 60%, 20%,

and 20% for the training, validation, and testing sets, respectively. A complete description and statistics of each dataset are given in the following subsections.

4.1 Urdu Date Dataset

There are multiple standard ways to write dates in Urdu. Participants were given the freedom to write the dates in any format. Urdu follows the Gregorian calendar for writing dates. They used slash '/', hyphen '-'and dot '.' separate day, month and year. Some people drew a curvy line beneath the year (عیسوی ۲۰۰۷), which represents سن (san) in Urdu, and signifies "the year" in English. Some people used the hamza 'ء' at the end of a year, which represents عیسوی (Aiswee) in Urdu and signifies "A.D." in the Gregorian calendar. Some people also wrote the name of a month in Urdu letters (جولائی means July) instead of writing it in numeral format. From English influence, some people also used the dot to separate month, day and year. This is a challenge in date recognition since the dot is similar in shape with the Urdu digit zero. Some samples are shown in the Table 1. Of the 318 total date samples, 190 belong to the training set and 64 images to each of the validation and test sets.

Table 1. Different Samples of Urdu Dates

English Date	Urdu Date
12/2/2007	۱۲/۰۲/۲۰۰۷
7-7-2007	۷-۷-۲۰۰۷
7 July, 2007	۷ جولائی ۲۰۰۷
16-Jul-2007 A.C	۱۶ جولائی ۲۰۰۷ء
8.8.2007	۸۰۸۰ ۲۰۰۷

4.2 Indian Isolated Digit Dataset

Each form contains two samples of each isolated digit. In the numeral strings, each digit is repeated 14 to 18 times at different positions.

We performed segmentation of the numeral strings by using a segmentation algorithm to separate and extract the digits. After segmentation, a set of isolated digits was created. Segmentation was performed on the grayscale and binary images. All other datasets are available in true color except isolated digits. We extracted a total of 60329 where 33974 were selected for training, 13177 for validation, and 13178 for test. A sample of each extracted digit is shown in Table 2.

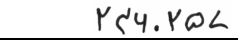
Table 2. Sample of Indian Isolated Digits

Nine	Eight	Seven	Six	Five	Four	Three	Two	One	Zero
۹	۸	۷	۶	۵	۴	۳	۲	۱	۰

4.3 Indian Numerical Strings Dataset

The data entry forms contain fields for 38 different numeral strings with varying lengths of 2,3,4,6 and 7 digits. Every digit, including the decimal point, is represented at least once in the strings. There are two ways in Urdu to write a decimal point; one way is to use dot (.) and the second is to use a hamza (ء). In most of the samples, the positions of the dot and hamza in numeral strings were located below the baseline. This could be a challenge in recognition of real strings as dot (.) looks like zero in Urdu. We divided this dataset into two sets – an integer set and a Real set. The latter includes decimal numbers of lengths 3 and 4. Samples for integer and real numeral strings are shown in Table 3 below. A total of 12914 strings were extracted which were divided into training (7750), validation (2584), and test (2580) sets.

Table 3. Sample of Different Numeral Strings

English (Numeral String)	Urdu (Numeral string)
47	
2460257	
1.50	
1.50	

4.4 Urdu Alphabet Dataset

The Urdu alphabet consists of 37 basic letters and some special characters (which consist of a combination of two letters). We added some of them as shown in Table 4. Similar to Arabic and Farsi [2] the words in Urdu are written in a cursive manner and the shape of a letter changes according to its position within a word. The data entry form includes one sample of each of the 37 Urdu isolated letters and each of the special letters. We have a total of 14890 letters with 8934 marked for training an 2978 for the validation and test sets each.

Table 4. Some Urdu Special Characters

Alif Mud AA	Bardi ye Hamza	Noon Guna	Chooti Ye Hamza	Hey Hamza
				

4.5 Urdu Words Dataset

We selected 57 Urdu (mostly financial related) words for our word dataset. This dataset could help in recognizing new and different challenges in the recognition of Urdu handwriting and word spotting. It includes words for weights, measurements, and

currencies. Samples of some of these Urdu words are shown in Table 5. The total number of images is 19432. These were divided into training (60%), validation (20%), and test (20%) sets.

Table 5. Sample of Urdu Words

Cash	Thousand	Cost	Decrease	Balance	Amount
نقد	ہزار	لاگت	کمی	بیلنس	رقم

4.6 Special Symbols Dataset

The data entry form also includes some special symbols that usually appear in any Urdu document. These symbols, which are also commonly used in other languages, are: comma (,), colon (:), at (@), slash (/), and pound (#). The total number of training samples is 1020. The validation set and the test set contain 340 and 345 images, respectively.

5 Ground Truth Data

For each folder that contains handwritten samples we have also included the ground truth data file. For each sample, the ground truth data file includes the following information: image name, content, number of connected components (CCs), writer number, length of a content, gender, and handwriting orientation. An example of the ground truth data for the numeral strings dataset is shown in Table 6.

Table 6: Examples of the Ground Truth data for Urdu Numeral String Dataset

Image Name	URD0110_P01_056.tif	URD0090_P01_029.tif
Content	0581294	47
Writer No.	URD0110	URD0090
Gender	F	M
Hand Orientation	R	R
Length	7	2
Length (No. of CCs)	7	2

6 Experiments and Error Analysis

6.1 Experiments

Recognition experiments have been conducted on the handwritten isolated Urdu digits. In image pre-processing, we did noise removal, grayscale normalization, size normalization, and binarization, etc. on all the grayscale images. In feature extraction, a 400D gradient feature based on Robert's operator [6] on each normalized image is extracted. In classification, a Support Vector Machine (SVM) using a Radial Base

Function (RBF) kernel function is applied. As a result, we have achieved an accuracy of 98.61% on the test set. The original Urdu Isolated numerals database includes the Training, the Validation, and the Test sets. Since we did nothing on the validation, we combine the Validation set to the Training set. Therefore, the number of samples in the Training set is 47, 151, and the number of the test samples is 13,178.

In image pre-processing, we did noise removal, grayscale normalization, size normalization, and binarization, etc. on all the grayscale images [4]. If the inputs were detected as binary images, they were converted to pseudo-grayscale images automatically. After removing the noise from a background-eliminated grayscale image, we cropped the image to remove the blank boundaries. Afterwards, we did size normalization using Moment Normalization (MN) [3] and grayscale normalization. Finally, we binarized the images based on the threshold calculated with the Otsu Method [5].

For extracting 32-direction gradient feature, the size of normalized image is set to 36 × 36 pixels. On each of 32 direction maps, 4 × 4 feature values are extracted in each block, and thus each image is divided to 9 × 9 blocks. Then, down sampling by Gaussian filtering is applied to reduce both the number of blocks and number of directions, and finally the feature dimensions is 400 (5 horizontal blocks × 5 vertical blocks × 16 directions) [6].

A Support Vector Machine (SVM) was chosen as a classifier for this research. An SVM constructs a separating hyperplane in the feature space, one which maximizes the margin between the data sets, and the kernel function chosen in this experiment is Radial Base Function (RBF). There are two parameters in RBF to be optimized: c and γ. c > 0 is the penalty parameter of the error term, and γ is the parameter in RBF. These two parameters can be are optimal chosen by the cross-validation. When lg(c) = 5 and lg(γ) = -7, the performance on training set achieve the highest recognition rate (97.60%). Thus, we set c = 32 & γ = 0.0078125, and then consider them as parameters in test.

6.2 Error Analysis

The recognition result has achieved accuracy with 98.61% (12995/13178) on the test set. ¾ errors can be grouped to four categories: (1) Almost ½ (88/183) errors occur among “2” (۲), “3” (۳), and “4” (۴) because of their similar topologies. In this category of errors, half of them occur in “3”, which are substituted by “2” since some “3” have small waves on their upper parts, shown in Figure 3. Moreover, some handwritten “2” (۲) and “4” (۴) look the same in Urdu, so these make the recognition more difficult. (2) As “6” and “9” are similar in Urdu as well, some errors (33/183) are caused by the ambiguous of “6” and “9”. All the substitution images are shown in Figure 4. (3) “5” and “0” in Urdu can have almost the same shapes which look like a circle, so the number of errors between these two classes is 12/183. Figure 5a shows all the errors between “6” and “9”. (4) Some short strokes in “0” misrecognized as “1”, and small “1” misrecognized as “0”. All are shown in Figure 5b. Other substitution images are caused by variations of some individual’s handwriting. For example, when the circles in “9” are very small (۹), they are misrecognized as “1”. The confusion matrix is shown below (Table 7).

Table 7. Results of Handwritten Urdu Isolated Numeral Recognition on Test set

Truth Label	Output									
	0	1	2	3	4	5	6	7	8	9
0	2009	3		1		**5**				
1	**7**	1548	1				1	1		1
2			1930	**12**	**13**					
3			**44**	1015	6					
4			**10**	3	1541					
5	**7**			1	1	952		1	1	
6	2	2	3			1	969			**19**
7	3			1	1	1		970		
8	2	1			2				902	1
9	2	**8**			1		**14**			1159

Fig. 3. Some substituted images in "3" which were mis-recognized as "2"

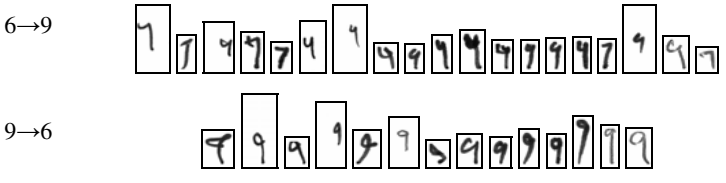


Fig. 4. All substitution images between "6" and "9"

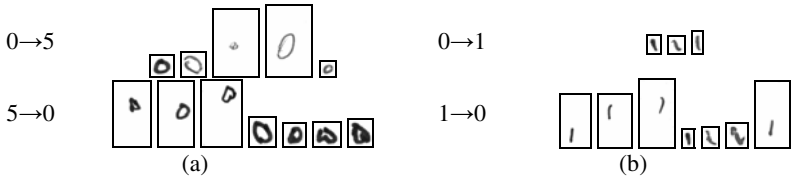


Fig. 5. All substitution images between "0" & "5" (a) and "0" & "1" (b)

7 Conclusion and Future Work

Although handwriting recognition has become a very popular area of Pattern Recognition for so many years, so far there exists no off-line Urdu handwriting database. Therefore, we created this comprehensive handwriting CENPARMI Urdu database which contains dates, isolated digits, numerical strings, isolated letters, a collection of

57 words, and a collection of special symbols. Experiments on the Urdu isolated digits have been conducted with a high recognition rate (98.61%). This database can be used for the future research in multiple purposes in Urdu documents analysis, such as Urdu word spotting, etc.

In the future, we will conduct more experiments on Urdu dates, numerical strings, isolated letters, words, and even symbols recognition. In addition, we will do further recognition on Urdu isolated digits. For example, we will analyze the errors especially in the four categories shown above to verify the recognition results or even reject the errors so that the recognition rate and/or reliability could be improved. We intend to make this database publicly available in the future.

References

1. Anwar, W., Wang, X., Wang. X.-L.: A survey of automatic Urdu language processing. In: Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, China, pp. 13–16 (2006)
2. Dehghan, M., Faez, K., Ahmadi, M., Shridhar, M.: Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM. Pattern Recognition 34(5), 1057–1065 (2001)
3. Liu, C.-L., Nakashima, K., Sako, H., Fujisawa, H.: Handwritten digit recognition: Investigation of normalization and feature extraction techniques. Pattern Recognition 37(2), 265–279 (2004)
4. Liu, C.-L., Suen, C.Y.: A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters. In: Proceedings of 11th International Conference on Frontiers in Handwriting Recognition (ICFHR), Montreal, Canada, pp. 278–283 (2008)
5. Otsu, N.: A threshold selection method from gray-level histogram. IEEE Trans. System Man Cybernet. 9, 1569–1576 (1979)
6. Shi, M., Fujisawa, Y., Wakabayashi, T., Kimura, F.: Handwritten numeral recognition using gradient and curvature of gray scale image. Pattern Recognition 35(10), 2051–2059 (2002)