Research Article

Development of Comprehensive Devnagari Numeral and Character Database for Offline Handwritten Character Recognition

Vikas J. Dongre and Vijay H. Mankar

Department of Electronics & Telecommunication Engineering, Government Polytechnic, Nagpur 440 001, India

Correspondence should be addressed to Vikas J. Dongre, dongrevj@yahoo.co.in

Received 5 April 2012; Accepted 31 May 2012

Academic Editor: Hyunchul Ahn

Copyright © 2012 V. J. Dongre and V. H. Mankar. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In handwritten character recognition, benchmark database plays an important role in evaluating the performance of various algorithms and the results obtained by various researchers. In Devnagari script, there is lack of such official benchmark. This paper focuses on the generation of offline benchmark database for Devnagari handwritten numerals and characters. The present work generated 5137 and 20305 isolated samples for numeral and character database, respectively, from 750 writers of all ages, sex, education, and profession. The offline sample images are stored in TIFF image format as it occupies less memory. Also, the data is presented in binary level so that memory requirement is further reduced. It will facilitate research on handwriting recognition of Devnagari script through free access to the researchers.

1. Introduction

With the advent of development in computational power, machine simulation of human reading has become a topic of serious research. Optical character recognition (OCR) and document processing have become the need of time with the popularization of desktop publishing and usage of internet. OCR involves recognition of characters from digitized images of optically scanned document pages. The characters thus recognized from document pages are coded with American Standard Code for Information Interchange (ASCII) or some other standard codes like UNICODE for storing in a file, which can further be edited like any other file created with some word processing software. A lot of research has been done in developed countries for English, European, and Chinese languages. But there is a dearth of need to carry out research in Indian languages. One common problem with the research is the need of benchmark database. To facilitate results on uniform data set, several document processing research groups have collected large numeral and character databases to make it available to the fellow researchers around the world. However, such existing databases are available only in few languages such as English, Japanese, and Chinese [1]. These standard databases include MNIST, CEDAR [2], and CENPARMI in English. Some work is also done for Indic scripts such as Bangla [3], Kannada [4], and Devnagari [5–8]. India is a multilingual and multiscipt country having more than 1.2 billion population with 22 constitutional languages and 10 different scripts. Devnagari is the most popular script in India. Hindi, the national language of India which is spoken by more than 500 million population worldwide, is written in the Devnagari script. Moreover, Hindi is the third most popular language in the world [9]. Devnagari is also used for writing Marathi, Sanskrit, Konkani, and Nepali languages.

In a developing country and emergingsuperpower like India, there is a need for the research and development of its own language technologies. The Department of Information Technology, Government of India, started a program on technology development for Indian languages [10] where language aspects are studied and developed. Another government undertaking Centre for Development of Advanced Computing [11] is actively involved in development of Indian languages fonts, translators. As a result of such initiatives, various research works for automatic recognition of printed/handwritten characters of various Indic scripts are
in progress. Some pioneering works on printed Indian scripts include [4, 12] for Bangla, [13] for Kannada, and [14] for Devnagari optical character recognition systems. There exist few studies on handwritten characters of some Indian scripts which include [15–20] for Devnagari characters. Research reviews on Devnagari character recognition are also available which includes [9, 21, 22].

Studies are reported on the basis of different databases collected either in laboratory environment or from smaller groups of the concerned population. The effective research work on handwriting recognition for Indic scripts is seriously hampered because of the unavailability of standard/benchmark databases, and those may be used for testing of algorithms and for comparison of results [3].

This paper describes an attempt for generation of a comprehensive database for handwritten Devnagari numerals and characters. This database has been developed with the view to make it available freely to the researcher community as a benchmark database for handwriting recognition research. The printed form of Devnagari numerals, vowels, and consonants are shown in Figures 4, 5, and 6. Sample handwritten form containing numerals and characters collected from a writer is shown in Figure 1. The present paper is organized as follows. Section 2 describes the details of offline database generation. Section 3 discusses statistical analysis of this work. Conclusion and further work direction are discussed in Section 4.

2. Devnagari Offline Database Generation Details

2.1. Data Collection. A sample A4 size sheet having blank boxes was designed. Persons of various ages, sex, education, and occupation were requested to write Devnagari numbers and characters. The only imposed restriction was that the character or numeral stroke should not touch the boundary of the boxes on the sheet and the vertical line made in the first box of every row. No restriction was imposed regarding colour of ink, thickness of lines, sequence of characters, and type of pen like ball pen or ink/gel pen. In case pen was not available with the writer, it was supplied at random from a set of different types of pens. The data was collected from 750 writers which included students of schools and colleges, office staff, workers, housewives, and senior citizens. The writers were carefully chosen to make the database representative. Persons of various languages and educational background like Marathi, Hindi were involved for writing on the blank sheets. Data was also collected from persons waiting in railway reservation centers and hospitals, which consists the mixture of all the categories mentioned earlier. Option of disclosing personal information was left to the writers so as to keep them free from stress that writing must be legitimate. Figure 1 shows sample handwritten data written by a writer.

2.2. Data Preparation. The A4 size paper sheet having the data written by various writers (Figure 1(b)) is digitized using Canon Canoscan Lide 100 flatbed scanner at 300 dpi. The images were stored in JPG format. It is cumbersome and time-consuming task to separate isolated symbols from the scanned image. Hence, various software modules were developed in Matlab to perform this task. The overall procedure is explained in the following. In all of the 750 samples, sheets are used in this work. Scanned images of the original paper sheets are also preserved in the original form for future use.

(1) Gray scale image is converted to binary for simplicity. In pattern recognition, we are concerned with shape and size of the object and not the color or gray level details. This also reduces data storage requirements as well as computation time.

(2) Isolated pixels (noise) are removed.

(3) The boundaries around the numerals and characters are removed using simple logic that it is the first and biggest continuous object. Other isolated groups of pixel are considered as desired data.

(4) Various rows are segmented using horizontal histogram approach [23]. Zero pixels in the histogram indicate separation of various rows. Each row is separately processed. Each row begins with vertical line as first object, which is ignored. This is used specially for preserving the dot present as a part of character ओ in Devnagari script. Otherwise, this character resembles with ए and all the images of ओ are lost.

(5) Useful characters segmented are stored in individual files. TIFF format is used for this purpose.

(6) The separated symbols are visually checked for proper shapes before sorting and storing in proper folders. 60 folders are formed for storing 10 numeral
databases and 50 character databases. A few samples of isolated numerals and characters from the present database are shown in Figure 2.

(7) Various image symbol files are serially numbered for further convenient use. Figure 7 shows size of numeral database for each numeral, and Figure 8 shows size of database for each character.

3. Statistics of Data Generated

Some Devnagari compound characters are not widely used in modern writing (e.g., ञ and घ). Some characters are written in more than one way, for example, ङ as ङ, ङ as ङ, and ङ as ङ. The database mostly contains first form of the numeral as it is written by most of the writers. Second form of the character is also written by few writers which is preserved in the database. The researcher may separate such data as per his/her need.

The ideal Devnagari script consists of curves and connected lines. Lines are not isolated from main symbol. But in practice, the handwritten documents and the number of strokes are unintentionally isolated due to inaccurate writing of writers. This imposes serious problems in document segmentation and further recognition. In the character segmentation stage, isolated strokes of modifiers are mistakenly considered as individual symbol and thus stored separately. Correctly segmented numerals and characters are shown in Figure 2. Isolated strokes and symbols in the handwritten document are shown in Figure 3(a). These captured strokes are rejected after visual inspection and removed from database. Also, ambiguous numerals or characters which may belong to more than one category are removed from database. Figure 3(b) shows such possible characters. Various characters are containing open curves and lines. Such characters cannot be uniquely categorized. Hence, they are also rejected. Some characters are improperly written by writers.
<table>
<thead>
<tr>
<th>Database no.</th>
<th>Symbol</th>
<th>Frequency</th>
<th>Database no.</th>
<th>Symbol</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$0$</td>
<td>490</td>
<td>5</td>
<td>$5$</td>
<td>482</td>
</tr>
<tr>
<td>1</td>
<td>$9$</td>
<td>492</td>
<td>6</td>
<td>$6$</td>
<td>522</td>
</tr>
<tr>
<td>2</td>
<td>$2$</td>
<td>634</td>
<td>7</td>
<td>$7$</td>
<td>458</td>
</tr>
<tr>
<td>3</td>
<td>$3$</td>
<td>582</td>
<td>8</td>
<td>$8$</td>
<td>482</td>
</tr>
<tr>
<td>4</td>
<td>$4$</td>
<td>501</td>
<td>9</td>
<td>$9$</td>
<td>494</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>5137</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 7: Devnagari Numeral database.

Such characters are also rejected In all of the 750 samples, sheets containing all the symbols were processed. Due to the reasons mentioned in the previous paragraph, various databases differ in frequency as shown in Figures 7 and 8.

It can be easily seen that the symbols having the combination of open curve and line (e.g., $\text{ा, अ, ड, ब, त, ज, न, ठ, ड, ध, श, ष, च, स, ह, घ, घं, घः, and घँ}$) have more chances of ambiguity and incorrectness. Such wrong strokes and ambiguous characters are removed from the final database. It may be noted that recognition efficiency for the previously mentioned characters may be poor.

Some characters got wrongly segmented as another valid character due to limitation of segmentation algorithm, for example, $\text{र}$ as $\text{र}$, $\text{आ}$ as $\text{र}$, and $\text{अ}$ as $\text{अ}$. Hence, it can be observed from Figure 7 that the frequency of numeral $\text{र}$ is more than that of other numerals. On the contrary, frequency of character $\text{र}$ is reduced (see Figure 8). It can also be observed from Figure 8 that the frequency of $\text{अ}$ is more (878) than that of any other character whereas frequency of $\text{अ, आ}$ are much less (195 and 92, resp.). It may be noteworthy that the frequency for $\text{अ}$ is even more than that of actual datasets scanned (750 images).

Some characters like $\text{क, ख, ज, and घ}$ are rarely used in modern writing. Hence, many writers skipped writing these characters in the blank datasheet provided. So, the frequency for previous characters is very low.

The character $\text{झ}$ is not a part of Devnagari database, rather it is a part of Marathi language which uses Devnagari script. The database for this character is also developed so that it may be useful for research on recognition of Marathi language.

Thus the quantity of numerals and characters in each category of database is reduced and varies as seen from Figures 7 and 8. It can also be observed that the symbol rejection rate is low for numerals than for characters. Hence, numeral recognition efficiency will be much better than character recognition efficiency.

### 4. Conclusion and Future Work

In this paper, we have generated a comprehensive database for Devnagari numerals and characters. Database of 5137 symbols is generated for numerals, and database of 20305 symbols is generated for characters. It is found that some symbols obtained need to be rejected as the writings of many persons are not recognizable by visual inspection. It will be impossible for computer software to recognize such symbols. The data images are stored in binary level and TIFF format for efficient storage and computational needs. This database will be further grown with more samples from variety of writers. Also, the database will be categorized as training set and test set randomly in near future. This database will be made freely available on http://code.google.com/p/devnagari-database/. This will surely help the research community for benchmarking their research results.

### Acknowledgments

The authors would like to thank Mrs. Rupali Dongre, Mr. Jitendra Bangari, and Mr. Prashant Kelzare for helping in digitization and sorting of the database. They would
also like to thank all the writers who contributed in this database.

References


