Amalgamated Approach for Devanagari Script Corpus for OCR & Demographic Purpose and XML for Linguistic Annotation

Conference Paper · December 2017

DOI: 10.1109/SITIS.2017.50

Maninder Singh Nehra
Malaviya National Institute of Technology Jaipur
14 PUBLICATIONS **121** CITATIONS
SEE PROFILE

Neeta Nain
Malaviya National Institute of Technology Jaipur
125 PUBLICATIONS **905** CITATIONS
SEE PROFILE

Mushtaq Ahmed
Malaviya National Institute of Technology Jaipur
66 PUBLICATIONS **399** CITATIONS
SEE PROFILE

Prakash Choudhary
Central University of Rajasthan
51 PUBLICATIONS **568** CITATIONS
SEE PROFILE

Amalgamated Approach for Devanagari Script Corpus for OCR & Demographic Purpose and XML for Linguistic Annotation

Maninder Singh Nehra, Neeta Nain, Mushtaq Ahmed , Prakash Choudhary, and Deepa Modi
Malaviya National Institute of Technology Jaipur, India-302017

ABSTRACT

In this paper, we present compilation of Hindi handwritten text image Corpus and its linguistics perspective in the field of OCR and information retrieval from handwritten document. Devnagari script is little bit complicated to enter a single character; it requires a combination of multiples, due to use of modifier. A mixed approach is proposed and demonstrated for Hindi Corpus for OCR and Demographic data collection. Demographic part of database could be used to train a system to fetch the data automatically, which will be helpful to simplify existing manual data-processing task involved in the field of data collection such as input forms like AADHAR, driving license, Railway Reservation etc. This would increase the participation of Hindi language community in understanding and taking benefit of the government schemes. To make availability and applicability of database in a vast area of corpus linguistics, we propose a methodology for data collection, mark-up, digital transcription, and XML metadata information for benchmarking and ZipF' s law to analyze the distribution and behavior of words in the corpus.

Keywords: Annotation, Corpus, Demographic, Hand-written, Hindi, OCR.

1. INTRODUCTION

A systematically designed corpus with advanced computer based technology provides a platform for wide number of applications in the field of human computer interaction with automation of handwritten document processing. The new era of corpus linguistic provides possibilities to explore all the diversity of language on the same grid at one place such as comparison and evaluation of various OCR algorithms and techniques, morphology, grammatical information, style of writing etc. Hindi is the national and official language of India. In the constitution of India 22 scheduled languages and Hindi is one of them. In India, there are 422 million Hindi speakers, it's about 41% of the total population of India. Hindi-Urdu speaking community is the fourth largest in the world after Mandarin Chinese, English and Spanish. In comparison to other languages, Hindi has a very less number of resources available in electronic data format. Most of the data available is in printed format. Data entry in Hindi needs more effort due to modifiers as compared to English. Many people live in rural areas where the data is collected through hand-filled forms. A huge amount of time, money and manpower is consumed in data processing and data entry every year for

978-1-5386-4283-2/17 $31.00 © 2017 IEEE
DOI 10.1109/SITIS.2017.50

successful implementation of various government schemes. If we could automatically scan printed Hindi text and convert it into its electronic text, we could bypass the tedious task of data entry easily. For this we need efficient handwritten text recognition systems. To train such systems we need a huge dataset. Therefore, availability of a standard dataset for Hindi handwritten text recognition is very important and necessary. This paper proposes and demonstrates a methodology to develop Hindi Corpus of full length handwritten text sentences having large volume of syntactic variations along with demographic information. The Corpus is structurally marked-up for para, line and words using the coordinates of corresponding segmented image. This structural mark-up is required for benchmarking of handwritten text segmentation algorithms. The Corpus also provides a huge repository of Hindi Ligatures, formed by a combination of basic vowels and consonants. The structure shows the handwritten image and transcription information of image at side panel on same view port. The ground truth and benchmarking of database, XML file is generated based on the database entries including all the meta-information of handwritten and demographic nature.

2. BACKGROUND

Handwritten document processing is becoming an important area in the field of pattern recognition. The availability of standard database of large scale documents is the main requirement for training and recognition, especially for off-line. A large number of handwritten text image databases have been developed by linguistic researcher's community. Some of the most widely used hand-written databases/dataset of different script have been identified for hand-written text recognition systems. Umapada Pal[1] developed hand-written city name database for Tamil. David Fern[2] developed BH2M a Barcelona historical marriage database. Li Deng[3] developed NIST modified digit dataset known as MNIST. U.pal[4] developed database for Devanagari legal amount words and Devanagari characters. ALAEI [5] introduced a PBOK database and ground truth for four different scripts: Persian, Bangla, Oriya and Kannada. The Chinese database CASIA[6] build by NLPR (National Laboratory of Pattern Recognition). [7] described the four level XML ground truthing of handwritten database. [8] Described a methodology to design a structure to annotation Urdu handwritten text image, segmented lines and words. Sabri and Mohammad [9]developed offline hand written database of Arabic text known as KHAAT with the help of 1000 writers of different countries. CALAM (Cursive And Language Adaptive Methodologies)[10] Urdu handwritten database.

It has been found from literature survey, that there is no standard handwritten tagged corpus or data set for Hindi language.

3. METHODOLOGY FOR DATABASE DESIGN AND STATISTICS

The proposed methodology of design a unified database for OCR and automation of handwritten document processing starts from a raw data collection according to 7 category & 19 subcategories and ends with completely mark-up XML file. To capture maximum diverse words, we have collected data from different categories such as history, architecture, science, arts, news economy, literature etc. The writers are free to replicate the machine printed text in their own handwriting.

The statistics of the Hindi corpus is as follows:

a) Handwritten form has a unique identification number which is the combination of category and subcategory. The same identification number is extended by hyphen for segmented lines and words.

b) To maintain the consistency of the database, the corresponding handwritten scanned image and its XML file also have the same unique identification number.

c) To cater maximum number of words in the database, the data collection is done in 7 different categories which is further divided into 19 subcategories.

d) 8-bit indexing was used for generating the form identification number; therefore a maximum of 256 forms can be stored in each subcategory. Thus, the database will consist of a total number of 3584 handwritten forms.

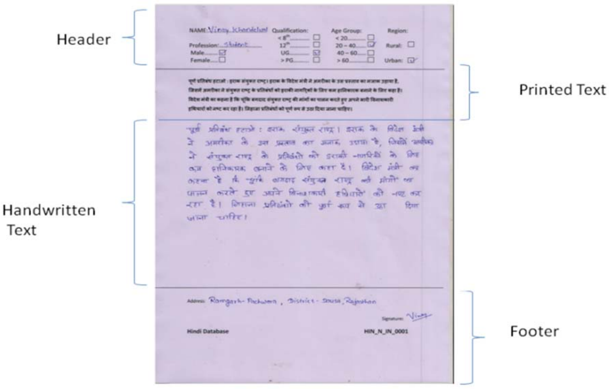
e) Each form containing 50 to 60 words. We would be able to provide a dataset of approximately 2,15,040 different handwritten Hindi words.

f) To achieve the maximum syntactic variations, forms will be filled up by different writers having diverse educational background and belonging to different geographical locations.

g) A hierarchical multi-level indexing is used to mark-up the image, segmented lines, words and ligatures with the coordinate's information and its digital transcription.

3.1 Handwritten Form Process Methodology

The handwritten form is divided into four sections for convenient segmentation of the data into machine printed text, handwritten text and demographic information of writer. It has been found from the filled handwritten forms that the writing style of the user changes a little bit when the writer fills the demographic information. The reason behind this is: because the writer frequently writes the demographic information usually in daily uses, so the fluency of writing the demographic information is different compared to his regular writing.


Header
Printed Text
Handwritten Text
Footer

NAME: Vinay Khandelwal	Qualification:	Age Group:	Region:
Profession: Student	< 8th □	< 20 □	
Male ☑	12th □	20 – 40 ☑	Rural: □
Female □	UG ☑	40 – 60 □	
	> PG □	> 60 □	Urban: ☑

[illegible]

[illegible]

Address: Ramgarh-Pachwara, District- Dausa, Rajasthan

Signature: Vinay

Hindi Database HIN_N_IN_0001

Figure 1. Designed handwritten form filled by writer.

The complete handwritten form of A4 size as shown in Fig. 1 is divided into four sections for different purpose of research as following :

Section 1: We tried to collect most of the demographic information such as: Name, Age, Gender, Education, Rural/Urban and Profession.

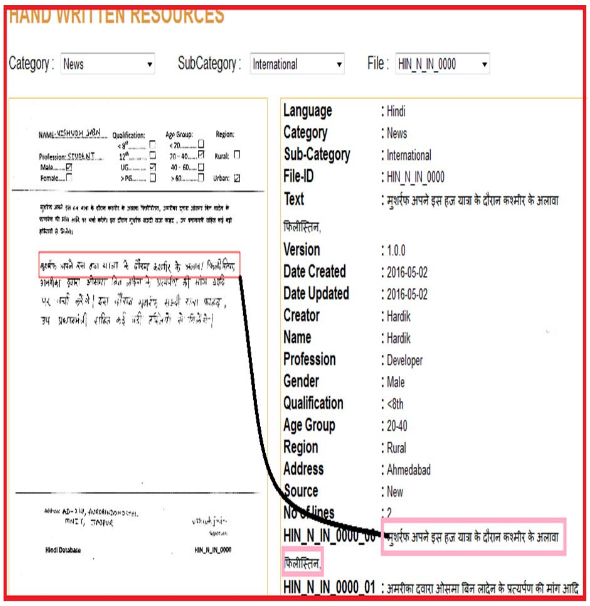
Section 2: Consists of a machine printed text of 3 to 5 lines.

Section 3: Blank to replicate machine printed text by writers in their natural handwriting. The syntactic variations in writing style writers was selected from distributed geographical region.

Section 4: Address, Signature, date of form filling and unique identification number of the handwritten scanned form as UID = HIN-N-IN-001 for form no.1.

3.2 XML File for Ground Truth

The ground truth annotation part has been done manually to achieve high accuracy rate in benchmarking as shown in Fig. 2, all the digitized transcription information along with the segmented coordinates of image, lines, words were stored in database. To reduce the manual data processing work in generating XML file a mapping has been done between the structure and database entries, which fetch all the information from manual data base entries and generate its XML file. That contains all the required information and digital transcription of the handwritten OCR and demographic information.



Category : News ▾ SubCategory : International ▾ File : HIN_N_IN_0000 ▾

Language	: Hindi
Category	: News
Sub-Category	: International
File-ID	: HIN_N_IN_0000
Text	: मुशर्रफ अपने इस हज यात्रा के दौरान कश्मीर के अलावा फिलीस्तिन,
Version	: 1.0.0
Date Created	: 2016-05-02
Date Updated	: 2016-05-02
Creator	: Hardik
Name	: Hardik
Profession	: Developer
Gender	: Male
Qualification	: <8th
Age Group	: 20-40
Region	: Rural
Address	: Ahmedabad
Source	: New
No of lines	: 2
HIN_N_IN_0000_00	: मुशर्रफ अपने इस हज यात्रा के दौरान कश्मीर के अलावा फिलीस्तिन,
HIN_N_IN_0000_01	: अमरीका दवारा ओसमा बिन लादेन के प्रत्यर्पण की मांग आदि

NAME: VISHUDH JAIN Qualification: <8th □ 12th □ UG ☑ >PG □ Age Group: <20 □ 20-40 ☑ 40-60 □ >60 □ Region: Rural: □ Urban: ☑
Profession: STUDENT Male ☑ Female □

Hindi Database HIN_N_IN_0000

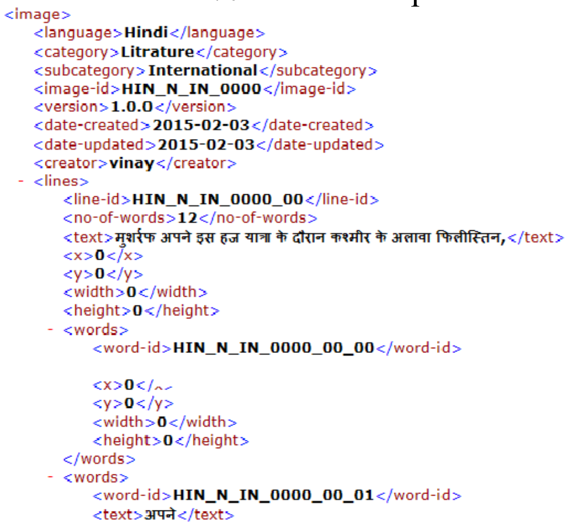
Figure 2. A snapshot of user interface structure displaying handwritten image, and bounding box of line coordinates information.

A snapshot of layered XML file for ground truth data is shown in Fig. 3, the XML file generates output in layered format (for image, line, word and ligature). All the tag sets has been verified manually with the input file to validate the XML file.

3.3 Statistics of the Database

The dataset contains more than 1600 handwritten text form, each were filled by different writers of different age groups and educational qualifications, out of which 75% writers were younger than 26 years and 25% were graduate students. Text pages are written by both males and females.

Information about name, age and address was collected on each page. Writers were asked to write forms in unconstrained environment in their natural handwriting with different pen styles and inks. To capture the maximum variance in data collection the domain of data collection is divided into 7 categories and 19 subcategories, category wise data collection is shown in figure 4. The data follows gaussian distribution with mean of 300 forms. The 66.66% of data is captured under 1-standard deviation and 100% of the data is captured under 2-standard deviation. The geographical statistics of the corpus is shown in figure 5. The data follows Gaussian distribution with mean of 290 forms.



```
<image>
   <language>Hindi</language>
   <category>Litrature</category>
   <subcategory>International</subcategory>
   <image-id>HIN_N_IN_0000</image-id>
   <version>1.0.0</version>
   <date-created>2015-02-03</date-created>
   <date-updated>2015-02-03</date-updated>
   <creator>vinay</creator>
 - <lines>
      <line-id>HIN_N_IN_0000_00</line-id>
      <no-of-words>12</no-of-words>
      <text>मुशर्रफ अपने इस हज यात्रा के दौरान कश्मीर के अलावा फिलीस्तिन,</text>
      <x>0</x>
      <y>0</y>
      <width>0</width>
      <height>0</height>
    - <words>
         <word-id>HIN_N_IN_0000_00_00</word-id>

         <x>0</
         <y>0</y>
         <width>0</width>
         <height>0</height>
      </words>
    - <words>
         <word-id>HIN_N_IN_0000_00_01</word-id>
         <text>अपने</text>
```

Figure 3. A snapshot of XML file for ground truth data.

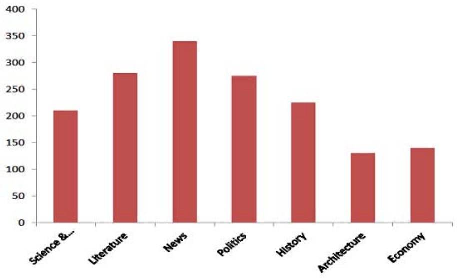


Figure 4. Category wise data collection

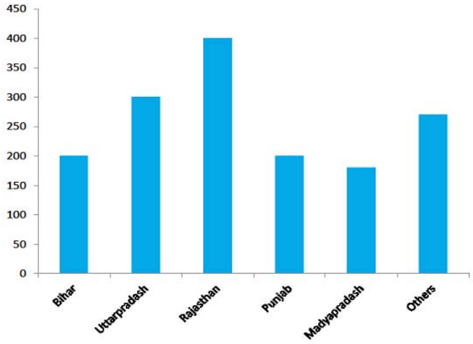


Figure 5. Geographically collection of data

4. FREQUENCY DISTRIBUTION OBSERVATION USING ZIP'F LAW

To analyze the distribution and behavior of words in a the corpus, we have apply ZipF law [11]. According that every natural language follows the law for frequency distribution of words. If f is the frequency of word in a corpus and r is the rank of word then:

$f \propto 1/ r$…………… (1)

ZipF's law states that, if words are arranged from the corpus in descending order of frequency, if w1 is the first most frequent word in the corpus, w2 is the second most frequent word, and so on. Then, the occurrence frequency of second word w2 is w12 times as the first word w1 and the third word w3 occurred roughly w13 one-third as often as the first word, and so on. From this it can be conclude that, the multiplication of the rank of a word r (rank one being the most frequent) by its frequency f, the product C would remain approximately the same for each word.

$wf_i = C/ wr_i$………… (2)

Frequency of words decreases very rapidly with rank. This can also be written as,

$wf_i = C(wr_i)^k$…………(3)

By taking log,

$\log (wf_i) = \log C + k \log (wr_i)$……4)

where k = −1 and C is a constant. So, log(f) and log(r) graph drawn between frequency and rank of a corpus must be linear with slop of −1.

5. CONCLUSION

A standard Devnagari script (Hindi) corpus including large scale of handwritten documents, XML and ZipF law to analyze the distribution and behavior of words in a corpus are explained. The aim is to make availability and applicability of a standard Hindi handwritten Corpus for wide range of pattern recognition application. A unified approach of collecting full length Hindi sentence and demographic information of writers could be used in a wide area of corpus linguistic applications such as: automatic information retrieval, signature verification, writer identification, benchmarking of various OCR algorithms and techniques etc. The structure uses hierarchical layers of annotation, which is further mapped between database and structure to create a single XML output which contains multiple layers of information.

REFERENCES

[1]. Umapada Pal, "Tamil Handwritten City Database Development",In In In

Proc. of ICDAR IEEE,pp. 793-797,2013.

[2]. David and Jon , “BH2M: the Barcelona Historical Handwritten Marriages database ”, In Proc. of ICPR, pp.256-261, 2014.

[3]. Li Deng “The MNIST Database of Handwritten Digital Image for Machine Learning Rearch”, In IEEE Magazine , pp.141-142,2012.

[4]. U. Pal, “Database Development and Recognition of Handwritten Devnagari Legal Amount Words”, In Proc. of ICDAR, pp.304-308, 2011.

[5] U. pal ,” Dataset and ground truth for handwritten text in four different scripts”, International Journal for pattern recognition and artificial intelgence,vol-26, issue-4,june 2012.

[6]. Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. 2011. CASIA Online and Offline Chinese Handwriting Databases. In Document Analysis and Recognition (ICDAR), 2011 International Conference on. 37–41.

[7]. Prakash, Nain and Nehra,”A framework for compilation of multilingual handwritten database:Four level XML ground truth”. SITIS-2015.

[8] P. Choudhary and N. Nain, “An Annotated Urdu Corpus of Handwritten Text Image and Benchmarking of Corpus”, MIPRO 2014, 26-30 May 2014.

[9]. Sabri and Mahmouda, “ KHATT: Arabic Offline Handwritten Text Database”, In Proc. of ICFHR,PP. 449 -455, 2012.

[10]. Choudhary, Prakash & Nain, Neeta , “ A Four-Tier Annotated Urdu Handwritten Text Image Dataset for Multidisciplinary Research on Urdu Script”, ACM Transactions on Asian and Low-Resource Language Information Processing. 15. 1-23. 10.1145/2857053, (2016).

[11]. StevenT. Piantadosi.,“Zipf's word frequency law in natural language: A critical review and future directions”, Psychonomic Bulletin Review 21, 5 (2014), 1112–1130.