ISSN (online): 2250-141X
Vol. 1 Issue 1, November 2011

A Survey of Nandinagari Manuscript Recognition System

Prathima G, Guruprasad K S Rao

prathimaguru18@gmail.com, ramaguruprasad@gmail.com

Abstract

Humans still outperform the latest Personal Computers in routine functions such as vision, reading and knowledge acquisition. Machine simulation of human reading has been the subject of intensive research from past several decades. Various techniques in the field of image and pattern recognition have been evolved but yet it is far from reality. Scanned document pose significant challenge to recognition of content. Innumerable volumes of such images although stored or archived well as scanned information cannot be reused or shared if they are not interpreted. This is still a research area to recognize the digitized image words in a scanned document and interpret these words to a fair accuracy. It becomes more challenging in case of Nandinagari scanned handwritten document when they need to be interpreted. While an ocean of wisdom lies there it is either not interpreted well and transformed to searchable text manually or is un-interpreted due to lack of knowledge and other external constraints that individuals face in today's world. This paper intends to provide an extensive survey to recognize the Nandinagari scripts and its interpretation. It also makes an attempt to propose a model System Architecture for processing Nandinagari images using Optical Character recognition.

1. Introduction

Very few scholars today can read and interpret the Nandinagari scripts. The life span of this script is over a half decade (from 7^{th} century to 15^{th} century) and un measurable information will be lost for the next generation if it is not preserved in a proper format.

One could infer that Nandināgarī, Nāgarī and Devanāgarī are very close and show only minor distinctions. From a study of the available evidence in manuscripts, however, it is clear that the scripts are indeed related as sister scripts, but that there are significant and systematic differences which justify us in considering them as different scripts. In the case of Nandināgarī the separate status is very clear but because of neglect and misleading statements in secondary literature it is frequently noted that scholars who try to read the script without proper preparation fail and have to give up.[1]

If we can automate this process a lot of man power and human effort could be saved. It is very useful to the society as Nandinagari/ Devanagari scripts are used by millions of users across the globe and especially in India more than 500 million people use Devanagari script for documentation. Also it is proved that these scripts are very efficient language for processing the information using computer system. In near future this will become a common language for e-processing. The Scope of this paper is hence unlimited.

ISSN (online): 2250-141X
Vol. 1 Issue 1, November 2011

An extensive literature survey reveals that this is the first attempt to recognize the Nandinagari scripts and its interpretation. Even in handwritten Devanagari very few research reports are published recently. This is challenging because of complex character set and Nandinagari is the earlier version of Devanagari.

The name of the script "Nagari" is supposed to be derived from "Nagara" a title of Pataliputra (Patna), the present capital of Bihar. But it might have been devised by the merchant community or "Nagarakas" and attained this name, due to its association with them. Nandi Nagari and Deva Nagari are two types of the Nagari script. Nandi Nagari script was used to write Sanskrit language, and most of the Sanskrit copper plate inscriptions of the Vijayanagar period are written in that script

Manuscripts in Nāgarī, Nandināgarī, and Devanāgarī, which were prevalent in the North, South and Central part of the Indian subcontinent respectively, have been intensively used and studied in modern indological studies of Sanskrit texts for at least two centuries. It may therefore come as a surprise that till now the distinctive features of these scripts have hardly been analyzed, and that their mutual relations and development from early stages in the second half of the first millennium till their pre-modern and modern forms have not yet been systematically and comprehensively studied. Such a study is very promising in several respects, especially for Indian manuscript ology. It can be expected that such a study, if conducted with sufficient thoroughness, will enable us, for instance, to assign a relative chronological place to a manuscript on the basis of the calligraphic style of the script.

Nandinagari script is the western variety of the archaic Nagari script of northern India and Nagari is also found in the inscriptions and manuscripts available in the western part of a few southern states; for example, south Maharashtra, Karnataka and Andhra Pradesh. That is why Nandinagari is also known as southern variety of Nagari.[2]

Origin Nandinagari is a descendent, as all indigenous scripts of India and Southeast Asia are, of the Brahmi script. This script was developed through various stages. It is closely related to northern Nagari which took its identifiable shape as early as the tenth century A.D. The modern Devanagari, which is now used for writing and printing Sanskrit, Hindi, Nepali, Rajasthani and Marathi, is a refined and standardized form of old or archaic Nagari script. Most probably, since the refined Nagari is used for writing Sanskrit which is venerated as devabhasha (language of divinities), it is called 'Devanagari'.

Nandinagari has never been used for printing and hence it lacks the necessary refinement and standardization. Nevertheless, its importance in the areas of epigraphy can't be ignored. There are innumerable manuscripts written in Nandinagari, covering vast areas of knowledge, such as Vedas, philosophy, religion, science and arts. These are preserved in the manuscript libraries, particularly those in the southern regions of the country.

2. Meaning and epigraphy

It is difficult to present any exact etymological meaning of the name 'Nandinagari'. The first part of the term 'Nandi' is rather ambiguous in the present context. It may mean 'sacred' or 'auspicious' (cf. Nandi verses in Sanskrit drama). Nandi is the name of Lord Siva's

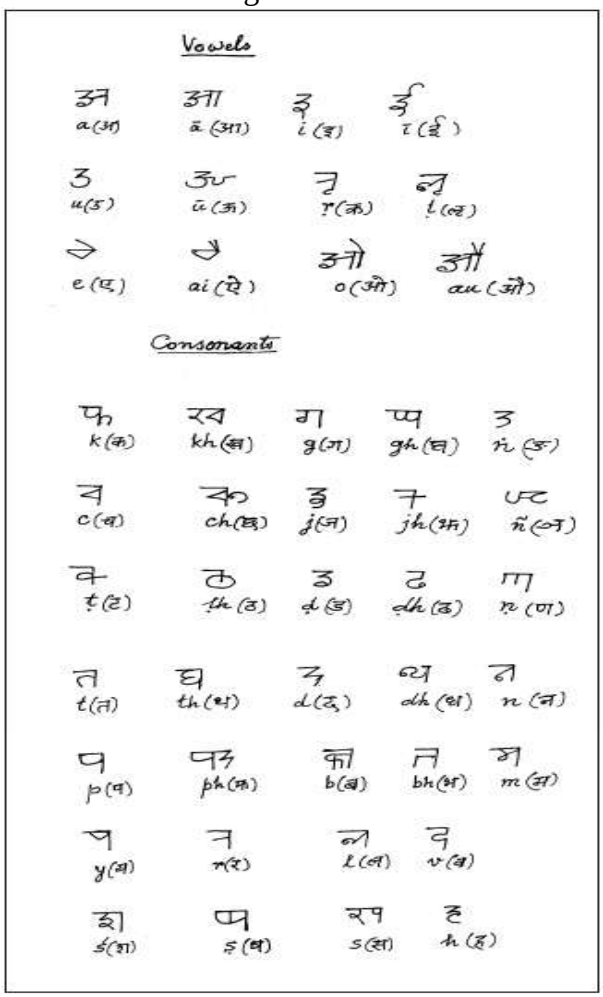
ISSN (online): 2250-141X
Vol. 1 Issue 1, November 2011

brishavahana (bull vehicle). Nandi bull is widely worshipped in the South, particularly in Karnataka.

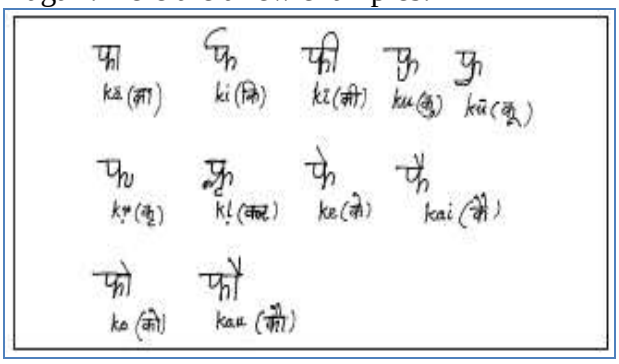
It may be mentioned that majority of the inscriptions, particularly the Sanskrit ones, of the period of Vijayanagara Empire are inscribed in Nandinagari. A. C. Burnell held that Nandinagari was used exclusively for writing on palm leaf. This view is supported by Shivaganesha Murthy also. But the existence of innumerable inscriptions in Nandinagari invalidates this view altogether. We can only say that, there developed two types of Nandinagari, slightly differing from each other - one used in the inscriptions, inscribed with chisel and the other used in palm leaf manuscripts written with the help of stylus. Obviously the latter type is rather cursive.

Some of the modern epigraphists opine that Nandinagari is less legible. But this view is also not correct. To one, who can read the Nagari of medieval inscriptions and manuscripts, Nandinagari is perfectly legible and transparent. It seems to be practical to furnish a chart of the basic letters of the Nandinagari alphabet before discussing the characteristics and variations of the script. The constituents and ligatures in conjunct consonants in Nandinagari are easily identifiable as they are in Devanagari. There are, however, a few exceptions. Though Nandinagari script is no longer in vogue, neither for printing nor for writing, no scholar of Sanskrit language and literature can afford to remain ignorant of this script. For the students of Indian epigraphy and paleography, learning Nandinagari is a must. *It is also proved to be very useful for those who are engaged in in-depth textual study of Virasaiva and MadhvaVaisnava works.* Nandinagari is helpful in another way: one who is proficient in it can read or learn Jain Nagari script with less effort.

The characters are given below:



The system of adding medial vowels in Nandinagari closely resembles that of Nagari. Here are a few examples:



ISSN (online): 2250-141X
Vol. 1 Issue 1, November 2011

3. Document image processing

Representation of documents as images is undesirable because of impossibilities of many common user level operations like editing, searching and large storage requirements. These limitations can be overcome by representing the content as text which takes very less space and is also convenient for processing. As an illustration a digitized document of A4 size may take 11 MB when represented as image in a database. The text format brings down the storage requirement for the whole document to a few KB and also makes them suitable for further processing.

Document classification is a problem in Information Science. This can be based on various FEATURES such as

- image level features,
- structural features or
- Classification based on textual features such as word frequency and word histogram or
- On structural information from tags.

A lot of advances have been made in the METHOD of classification such as:

- Clustering and pattern recognition methods.
- Strategies for modeling complex data and mining large data sets.
- Methods for the extraction of knowledge from data and
- Application of advanced methods in specific domain of practice.

Document classifications can also be based on TYPES:-

- Supervised document classification.
- Unsupervised document classification.
- Semi supervised document classification.

4. Proposed Model

We intend to take the following steps in the proposed model for Nandinagari (handwritten) recognition and its text interpretation as shown in Figure 1 :-

Step 1:- Obtain the digitized image of Nandinagari manuscript.

Step 2:- **Preprocessing:**

- Noise removal using filters
- Binarization, thinning of image etc.
- Cleaning techniques and filtering mechanisms.

Step 3:- **Segmentation**: This could be one of the following-

- Character Based
- Word Based
- Sentence based or line level
- Other techniques such as wavelet or curvlet

Following are some of the challenges faced in Nandinagari during this step:

1. The characters are touching
2. Joined handwritten letters

The characters should not be touching each other .Segmentation of touching characters has been one of the toughest jobs in text recognition and this alone has remained one of the toughest areas of research.

Step 4:- This step can be subdivided into following four sub divisions-

1) ***Feature Extraction*:** Feature extraction can be based on one of the following types-
   - Template based
   - Structural
   - Morphological
   - Statistical
   - Or any other optimal techniques.
2) ***Training***
3) ***Classification***
4) ***Matching techniques***

ISSN (online): 2250-141X
Vol. 1 Issue 1, November 2011

Step 5:- Character recognition and repeat the steps for the entire documents.

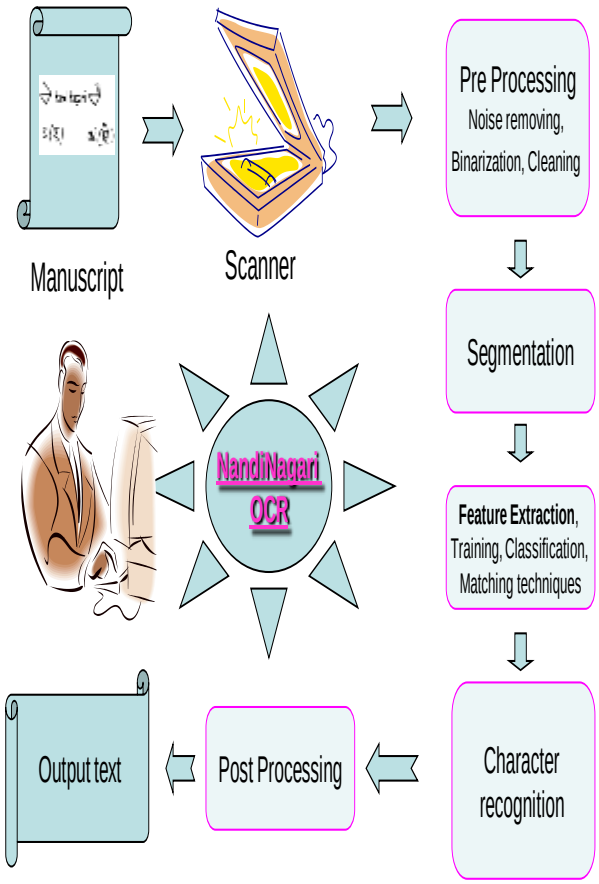


Figure 1: Nandinagari OCR model System Architecture

5. Optical Character Recognition:

OCR is used to recognize characters using connected components, segments, stroke analysis. Once all chars in the given word are recognized the word is compared against a vocabulary of potential answers for the final result. OCR translates the shapes and patterns of machine made characters into corresponding computer codes. Though most advanced systems are able to recognize multiple fonts they can process only standard fonts. Intelligent Recognition (ICR) recognizes both machine print and hand printed fonts. New generation technology based on neural networks is used to convert handwritten or machine printed data to ASCII text[3, 4].

OCR is a mechanical or electronic translation of scanned images of handwritten, typewritten or printed text into machine encoded text. OCR makes it possible to edit the text, search for a word or phrase, store it more compactly and apply techniques such as machine translation, text to speech, text mining to it.

OCR is a field of research in pattern recognition, Artificial Intelligence and computer vision. The accurate recognition of Latin script, type written text is now a largely solved problem. Other areas like handwriting and printed text in other scripts are still the subject of active research. Almost all Indic scripts are cursive in nature making them hard to recognize by machines. OCRs could be template based or feature based. Extracting the information is based on alphabetic structure such as recognizing the letter based on shape, contour, pattern recognition, stroke analysis, vertical/horizontal projections.

We need a method to access the content of these documents. This can be attributed to a number of challenges in the form of poor quality documents, complex nature of the script and relatively fewer years of research on Indic OCR [5,6,7]. Unfortunately, the progress in recognition technologies for Indic scripts such as Devanagari (used by Sanskrit, Hindi and other languages) has **not** been at par with the growth in the digital document collections.

ISSN (online): 2250-141X
Vol. 1 Issue 1, November 2011

6. Conclusion

1. Only handwritten scripts are available.
2. Character segmentation is more complex as a lot of variations in the content is available
3. If we consider all the variations, accuracy would be more. This is quite challenging.
4. Different forms of Nandinagari are available.
5. Not a real breakthrough is achieved in bringing Nandinagari to printed form for a specific font. Research is still in progress in many leading institutes
6. Nandinagari OCR is not available.
7. Nandinagari could come with/without proper shirorekha as in Devanagari scripts
8. Previous and next character to be considered while taking single character.
9. Identifying samyuktaksharas is challenging.
10. Sanskrit documents are available in more than 30 scripts such as Nandinagari, Nagari, Madhyanagari, Devanagari, Malayalam, Tamil, Telugu, Marathi, Manipuri etc. Hence one has to study the variations before preparation of the framework that hence in character recognition.
11. Info available anywhere can be fetched easily using this system and it is very useful.
12. Handwriting varies across scripts written by same people.
13. Handwriting varies across scripts written by different people.
14. Typical variations in this language could be over 13,000.
15. If a page contains 2000 characters, then by eliminating repeated characters, one can achieve 10%, processing faster than the original.

REFERENCES:

[1] 14th World Sanskrit Conference - Nandināgarī manuscripts: distinctive features, geographical and chronological range Saraju Rath International Institute for Asian Studies Leiden, The Netherlands. http://www.indology.bun.kyoto-u.ac.jp/14thWSC/programme/index.html

[2] National Mission for Manuscripts www.namami.org – Fourth and Seventh Annual report - Paleographical Importance of Nandinagari BY: SATKARI MUKHOPADHYAYA.

[3] A. Bhardwaj, S. Kompalli, S. Setlur and V. Govindaraju. An OCR based approach to word spotting in Devanagari documents. In Proceedings of the 15th SPIE - Document Recognition and Retrieval, volume 6815. 2008.

[4] N. R. Howe, T.M. Rath and R. Manmatha. Boosted decision trees for word recognition in handwritten document retrievals. In Proceedings of the SIGIR, pages 377-383, 2005.

[5] B. B. Chaudhuri and U. Pal. An OCR system to read two Indian language scripts: Bangla and devanagari (hindi). in Proc of ICDAR, pages 1011–1015, 1997

[6] C. V. Jawahar, M. N. S. S. K. Pavan Kumar, and S. S. Ravi Kiran. A bilingual OCR for hindi-telugu documents. Technical Report TR-CVIT-22, IIIT,Hyderabad, 2002.

[7] V. Bansal and R. M. K. Sinha. A devanagari OCR and a brief review of OCR research for Indian scripts. In Proc

ISSN (online): 2250-141X
Vol. 1 Issue 1, November 2011

of STRANS01, 2001.