A New Architecture for Brailee Transcription from Optically Recognized Indian Languages

Omar Khan Durrani K C Shet

Abstract

In order to bridge the digital divide between the sightless and sighted and to encourage literacy in them, we have designed an architecture for transcribing Braille from optically recognized Indian language. The system will help to convert masses of information in different Indian languages into a tactile reading form. The system mainly consists of OCR modules designed in an efficient manner to promote portability and scalability. In first section, we have introduced the importance and necessity of the work with successive sections clarifying briefly the properties of Braille and Indian scripts. We have also described the OCR work done with respect to Indian languages and the related work to our system. Finally, the System architecture is explained clearly followed by some conclusion and future work. The paper also identifies the needs to be fulfilled to percolate the benefit of the technology developed to the masses.

Keywords : Visually disabled, Digital divide, Braille Script, Indian Languages, Optical Character Recognition, Braille translation.

0. Introduction

Braille is a system of tactile reading and writing for visually disabled people. Each character or cell is made up of 6 (2*3) dot positions, 64 possible characters are available by using any one or a combination of dots. This system exists even today, 150 years after Louis Braille worked out its basics.

The sightless community in India and the developing world faces a tremendous hindrance in getting access to printed reading material which are scarce and communicating with the sighted community in writing, due to the difference in the script systems. Consequently, they are impaired in their educational opportunities as well as in the mainstream employment opportunities. The lack of readily available Braille material had restricted the literacy level to just three per cent of the visually impaired population. There was a pressing need for schools to generate Braille material through the computer, indigenously, to keep costs down, and in the various Indian languages to help the users. This also helps the visually impaired in teaching, learning, reading, writing and printing.

With technology making gigantic leaps and bounds, it is necessary for us to progress with this perception and try to fulfill the needs of forgotten sections of society, hence bridging the digital divide between various segments of the people. Keeping this as objective we have an system Architecture which takes input, the image file image preprocesses it and then by using the respective Optical Character Recognition (OCR) software to recognize the various composite characters and convert them into digitized text, which is then edited by a multilingual editor which includes facilities for Braille editing, Braille translation and multilingual screen reader module. The resultant Braille script can then be allowed for embossing with a suitable Braille printer.

Section 2 explains properties of Indian scripts, which are to be to taken care by the OCR engine for each Indian language. Section 3 illustrates the Braille script needed for Braille translation and Braille printing. Section 4 gives details about the work related to our system proposal. Section 5 describes the system architecture and Section 6 brings out the conclusion.

1. Properties of Indian Scripts

In India, there are eighteen official (Indian constitution accepted) languages, namely Assamese, Bangla, English, Gujarati, Hindi, Konkani, Kannada, Kashmiri, Malayalam, Marathi, Nepali, Oriya, Panjabi, Rajasthani, Sanskrit, Tamil, Telugu and Urdu... Twelve different scripts are used for writing these official languages. Examples of these scripts are shown in Fig. 1. Most Indian scripts originated from ancient Brahmi through various transformations [13]. Two or more of these languages may be written in one script. For example, Devnagari is used to write Hindi, Marathi, and Rajasthani, Sanskrit and Nepali languages, while Bangla script is used to write Assamese and Bangla (Bengali) languages. Apart from vowel and consonant characters, called basic characters, there are compound characters in most Indian script alphabet systems (except Tamil and Gurumukhi scripts), which are formed by combining two or more basic characters. The shape of a compound character is usually more complex than the constituent basic characters. In some languages, a vowel following a consonant may take a modified shape, depending on whether the vowel is placed to the left, right, top or bottom of the consonant. They are called modified characters. In general, there are about 300 character shapes in an Indian script [13].

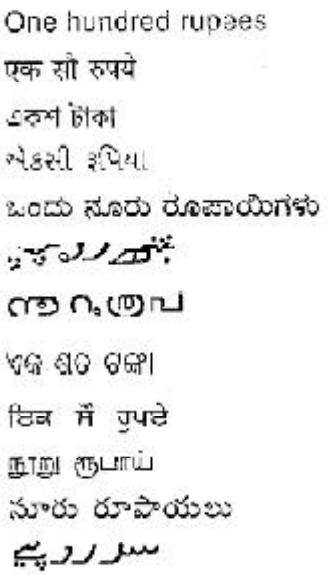


Fig. 1. Examples of 12 Indian scripts:

In some Indian script alphabets (like Devnagari, Bangla and Gurumukhi, etc.), it is noted that many characters have a horizontal line at the upper part. In Bangla, this line is called matra while in Devnagari it is called sirorekha. However, in this paper, we shall call it as head-line (see Fig. 2). When two or more characters sit side by side to form a word in the language, the headline portions touch one another and generate a big headline. Because of these, character segmentation from word for OCR is necessary. In some scripts, however, (like Gujarati, Oriya, etc.) the characters do not have headline.

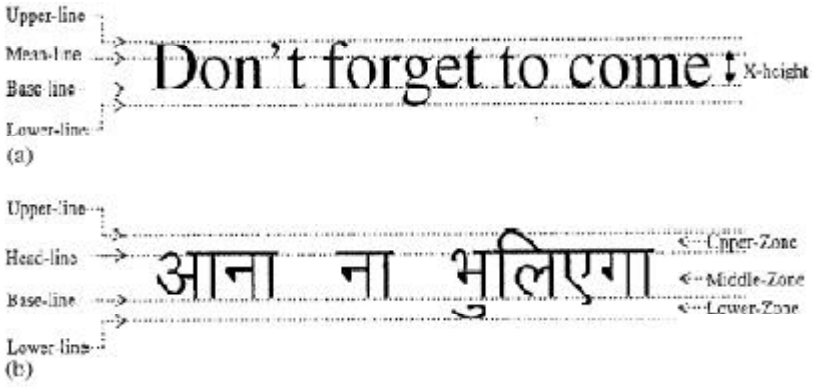


Fig. 2. Different zones of (a) English and (b) Devnagari text lines.

In most of the Indian languages, a text line may be partitioned into portion above the head-line, the middle-zone covers the portion of basic (and compound) characters below head-line and the lower-zone is the portion below base-line. Those text where script lines do not have headline, the mean-line separates upper- and middle-zone, while the base-line separates middle and lower-zone. An imaginary line, where most of the uppermost (lowermost) points of characters of a text line lie, is referred as mean-line (base-line). Examples of zoning are shown in Fig. 2. In this case, the head or mean-line along with base-line partition the text line into three zones.

2. Bharti Brailee

Bharti Braille is the standard prescribed in India for preparing Braille documents in all the Indian languages. The standard uses the six-dot system as in normal Braille but the cell assignments are corresponding to the aksharas of the Indian languages. Very simply, Braille is used as another script to text in all the Indian languages. This is possible on account of the phonetic nature of the languages where the writing system follow rules for displaying syllables rather than the basic vowels and consonants. The six dots Braille standard conforms to the following arrangement of dots, in three rows and two columns. The dots are numbered as indicated, below

1 o o 4

2 o o 5

3 o o 6

The six-dot system provides for displaying 64 different patterns. Of these, only 63 may be used for representing the aksharas. The 64th pattern is a cell without dots and is implied to represent the space character. In English Braille, the 63 different cells represent the letters of alphabet (26), ten punctuation marks, fourteen frequently used short letters and the rest assigned special meanings.

It may be noted that the assignment of meanings to each cell has no direct relationship to the set of displayable ASCII characters (96 in use). The meaning of a cell is to be interpreted in the context in which the cell is present, such as the cell preceding it, whether it appears in the beginning of a word, etc..

A sheet of printed Braille will have a series of cells embossed on thick paper and passing one's forefinger over each line of embossed cells can sense the embossing. A standard sheet of Braille has about forty cells per line and may contain 20 or more lines.

Bharti Braille is thus a system for writing syllables using a basic set of 63 shapes, each corresponding to a cell. Here, the most basic approach to writing syllables using generic consonants has been used. A syllable in Indian languages can take any one of the following forms.

(i) A pure vowel V. (ii) A pure consonant and a vowel CV. (iii) A conjunct with two or more consonants and a vowel CC..V.

Besides the syllables, special symbols are also used. These include modern punctuation marks as well.

In Bharti Braille, the basic vowels and consonants of the languages have been assigned individuals cells. Across the language of the country, between 13 and 18 vowels are in use and the consonants are between 33 and 37 in number. Thus more than 50 cells have been assigned for the basic vowels and consonants leaving the rest for special marks.

The cell assignment for a consonant assumes that the consonant has an implied vowel "a" as part of it. A pure consonant (also known as generic consonant) has no vowel and so to distinguish a basic consonant from its generic equivalent a special symbol is used in the writing systems. This is known as the halanth and its shape is a language specific ligature added to the shape of the basic consonant. Bharti Braille has set a part one cell for this purpose and this cell placed before the cell for a basic consonant turns it into generic consonant.

The idea here is that one can use this principle to write syllables in the CC..V form simply by concatenating the cells for each generic consonant.

The cell assignment corresponding to the basic vowels and consonants are similar to the assignments of the English alphabet where the sounds match. But only about 25 can be matched this way. Cells in standard Braille, which correspond to specific two letter contractions, have been chosen to take care of the aksharas such as the diphthongs and the aspirated consonants. In assigning the cells, a superset of the aksharas from all the Indian languages has been taken into consideration

Here (figure 3) are the assignments . It is also observed that some of the aksharas have been assigned two cells. The first of the two cells will invariably be a cell with just one dot, typically dot 5.

The understanding here is that the following cell has to be interpreted differently. Such schemes where a special symbol is employed to provide specific interpretation of the following character are common with computer systems and the special character is known as the escape character.

Bharti Braille confirms to the syllabic writing system followed for all the Indian languages and syllables are just written using the cells assigned for the consonants and vowels.

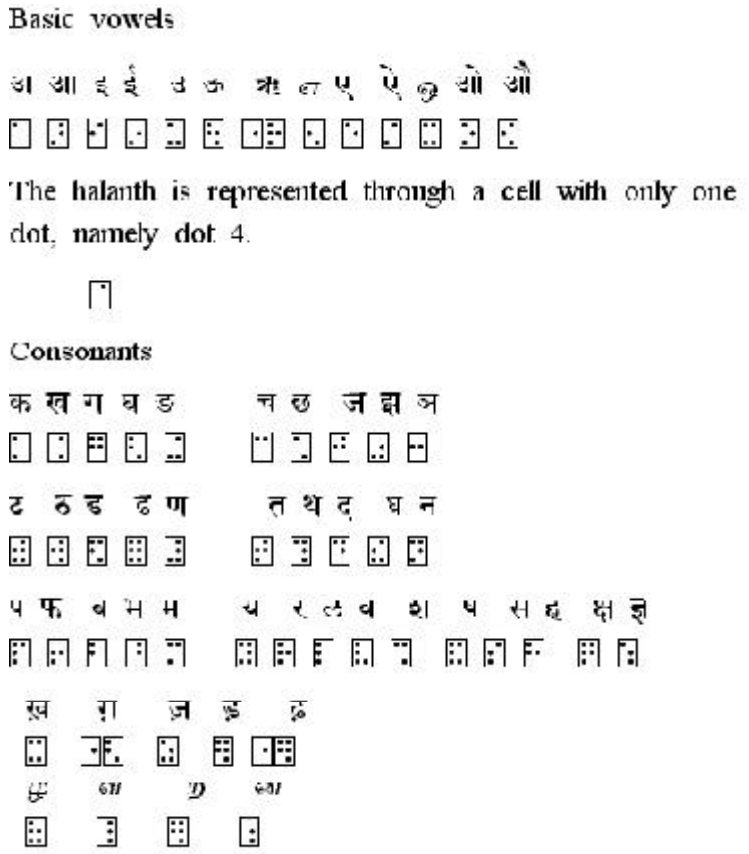
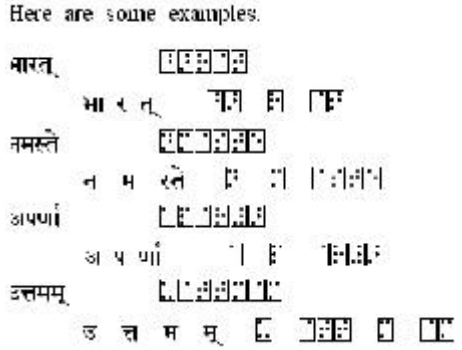


Figure 3

A pure vowel is always shown using the cell assigned. A basic consonant is always shown using the cell assigned for the consonant. A consonant vowel combination is shown using the respective cells. Normally a pure vowel will not follow a consonant and will appear only at the beginning of a word. However , there are many exception to this rule as explained in [4].



3. A view of OCR work on Indian Languages

In fact, there is not sufficient number of studies on Indian language character recognition. Most of the pieces of existing work are concerned about Devnagari and Bangla script characters, the two most popular languages in India. Some studies are reported on the recognition of other languages like Tamil, Telugu, Oriya, Kannada, Panjabi, Gujrathi, etc. [13]. Structural and topological features based tree classifier, and neural network classifiers are mainly used for the recognition of Indian scripts.

From the previous work done and experiences on various Indian script it is observed from [3,5,6,9,10,111,14,15] that the methods for preprocessing steps as well as method adopted for text and word separation can be common for all languages. Examples are Histogram based thresholding for binarization, Logical-smoothing approach for Noise cleaning, Hough transforms for skew detection, horizontal and vertical projection profiles for text and word separation are commonly used for Indian script. Structure and template based feature extraction and tree classification for Brahmi script (Devnagiri, Gurumukhi, Bangla etc.) are found to be better. For Dravidian script (Tamil, Telugu, Kannada and Malayalam) it varies due to some diverse nature among the scripts. Feature extraction using zoning and Simple Vector Machine (SVM), Nearest Neighbor (NN) classifications are found to be better (refer table 1).

4. Related Work

Galileo Reading System developed by the Robotron group has a scanner at the top and keypad at the front base. It is attached with speaker and floppy drive .It has an in-built hard disc, a serial port to get connected to the computer, printer etc. It is a multi-lingual machine. The machine operates by receiving commands through the keypad giving response through the speaker. The machine scans and stores the document as image in its buffer and recognizes it .The resultant image /text can then be either sent to hard disc, floppy disc, Embosser or to the computer. After recognition, it reads out the document till the end. Images or text can be copied from the floppy/hard disc/computer to the buffer for recognition and reading or vice versa. The recognized text can be sent to the translator after which we can have a Braille output through the embosser. As the Indian counter part, we have a system Drishti developed by LERC (Language Engineering Research Center) for Telugu and is currently being extended for other Indian scripts. Galileo is restricted to roman script; lacks built-in Braille translation from ASCII code and prices are exorbitant in the Indian context. Drishti has no built-in multi-lingual editing and text-to-speech facility as well as Braille translation.

5. System Architecture

The System has three main modules, the Preprocessing module, OCR engine and the Multi-lingual editor with Braille translation (See Figure 4b).

5.1 Preprocessing module

Along with Preprocessing we have added text and word separation in this module making it common for all the Indian languages. The methods for each step are selected as mentioned in the example of section 3. The input image under goes Binarisation, Noise cleaning, Skew detection and Skew correction, and gets segmented into lines and words. Binarisation is the process of converting the gray level image into a binary image with foreground as white and background as black. The skew may be caused while placing the paper on the scanner, or may be inherently present in the paper. Even with lot of care, some amount of skew is inevitable. After finding the skew angle; we need to correct the skew. Text and word separation involves breaking the text in the page to lines and words, which are required to identify the script.

5.2 Language Detection

The segmented text and words are then used by the language detector, which matches the language specific, shape characteristics shown below for Indian languages [2].

Bangla, Assamese - Triangular shapes with headline

Devanagari, Punjabi - Vertical Lines with headline

Gujrati - Vertical line with right bent at bottom but without headline

Oriya - Vertical line and inverted curves

Kannada - Horizontal and double bowl curves but no headline

Malayalam - Circular and no headline

Tamil - Multiple enclosed areas with some exceptional

Linear shape characters but no headline

Telugu - Circular with a tick mark sigh on top

This information (stored in the data base) is used to detect language regions, in the text region. With the help of all the information collated above, we can run different O.C.R algorithm (engine) for different languages.

5.3 The OCR engine

The OCR engine consists of four steps; character segmentation, feature extraction, classification and post processing. These steps differ for each Indian script. Therefore we have a separate OCR Engine for each Indian script (see figure 4 a).

Under character segmentation, individual words are processed to obtain the components for the scripts. For Devnagiri, Gurumukhi, Bangla etc segmentation involves the removal of sirorekha (the horizontal bar). This separates the constituent components from a word. For Telugu and Kannada, component extraction implies the separation of connected components. The individual connected components here are not distinct letters. They can also be modifiers. This results in splitting of the single word into many separated components. Once the sirorekha is removed, for Devnagiri like scripts, the top, middle and bottom zones are identified easily. Top zone gets automatically separated with the removal of sirorekha. Bottom zone is identified from the projections. Components in top and bottom zones for Hindi are part of vowel modifiers. Each of these components is then scaled to a standard size before feature extraction and classification.

Feature extraction captures the essential characteristics of the symbol to help in classification. It finds the amount of data by extracting relevant information, usually results in a vector of scalar values. Features are also normalized for distance measurements. Classification compares the feature vectors to the various models in the database and finds the closest match. The output after classification is transformed into a format like ASCII set or ISCII (Indian Script Code for Information Interchange).

5.4 Multi-lingual editor

The Multilingual-editor takes the input the ASCII or ISCII code and provides proof reading. It will also have utilities like screen reading and Braille translation. A provision for giving proof reading in Braille Script can also be provided. IIT, Chennai has also developed such software, which can be utilized in this module

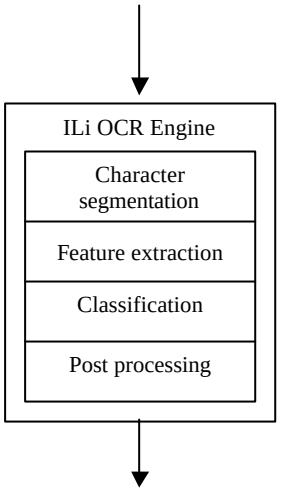


Figure 4 a : Expansion of OCR engine module.

Figure 4 b: System Architecture

6. Conclusions and Future Work

Using the methodology discussed above we can get better accuracy and output. This method has advantages like High efficiency, greater speed, less utilization of processing resources and less human intervention. With this system, Government can promote even higher education for the visually disabled. The work to bring this into a practical form has been started by us. We have already implemented the multilingual editor and extension to Braille translation is in progress hoping to finish in few months. Initially the work can be done for one or two fonts and later extended for multi font. Indian script OCR error correction module that corrects single character can be extended to post recognition error correction (spell check and morphological techniques with grapheme features of language script), which will improve OCR accuracy. A work in [5] has shown better results for Telugu and Hindi with SVM-PCA based OCR. Hence a study to adopt SVM-PCA for all Indian languages can be made keeping efficiency, accuracy and portability as objectives. Other aspects like Phonetic nature of the language and its reflection on the script, similarities and dissimilarities between script based on their origin and evolution, geometrical shape features for characterization of characters can also be addressed.

6.1 Acknowledgments

The authors wish to thank Prof. Anupam Basu of IIT, Kharagpur, Prof. Murali, PES, Mandya and Central Institute Indian Language, Mysore for suggestions and contributions.

Table 1: Different OCR systems on printed Indian script

Script	**System**	**Feature**	**Classification Technique**	**Accuracy claimed**
Devanagari	Pal and Chaudhuri [13]	Structural and template features	Tree classifier	96.5%
	Garain and Chaudhuri [13]	Run length-based template feature	Tree classifier	97.5%
Bangla	Chaudhuri and Pal [13]	Structural and template features	Tree classifier	96.8 %
Gurumukhi	Lehal and Singh [13]	Structural and topological features	Tree classifier	97.3%
Oriya	Chaudhuri et al.[13]	Structural and template features	Tree classifier	96.3%
Telugu	C.V Lakshmi and C Patvardhan [6]	Gradient based features	Symbol association information during segmentations	98%
Tamil	A G Ramakrishnan & Kaushik Mahata [10]	Geometric based moments	Based on spatial occupancy and Nearest neighbor	97%
Kannada	Ashwin and Sastry [11]	Zoning features	SVM classifier	97%
Bi-lingual Hindi-Telugu	C. V. Jawahar et al. [5]	Principle component transformation	SVM classifier	96%

7. References

1. Anupam Basu "Bharati Braille Information System: An Affordable Multilingual System for Empowering the Sightless Population", development by design, Bangalore, Dec 1-2, 2002.
2. Aditya Gokhale "Bi-lingual optical character recognition system For devanagari and english, Centre For Development Of Advanced Computing, GIST R&D Pune, Maharashtra, INDIA.
3. B B Chaudhuri, U Pal And M Mitra Automatic recognition of printed Oriya script Sadhana Vol. 27, Part 1, February 2002, pp. 23–34
4. Bharati Braille, a tutorial, http://acharya.iitm.ac.in/disabilities/index.html
5. C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran A Bilingual OCR for Hindi-Telugu Documents and its Applications Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
6. C. Vasantha Lakshmi and C. Patvardhan An optical character recognition system for printed Telugu text Pattern Anal Applic (2004) 7: 190–204.
7. Durre, K.P. (1990). "BrailleButler: A new approach to non-visual computer applications" Proceedings Third Annual IEEE Symposium on Computer-Based Medical Systems, University of North Carolina at Chapel Hill 1990. Los Alamitos, CA: IEEE Computer Society Press.
8. Directory of Aids and Appliances for the Visually Handicapped. Government of India, Rep. Ministry of Welfare, 1991.
9. G.S. Lehal, C. Singh, Feature extraction and classification for OCR of Gurmukhi script, Vivek 12 (1999) 2–12.
10. Kaushik Mahata and A. G. Ramakrishna, "A complete OCR for printed Tamil text ", Proc Tamil Internet 2001, Singapore, July 22-24, 2000, pp. 165-170.
11. T.V. Ashwin, P.S. Sastry, A font and size independent OCR system for printed Kannada documents using support vector machines, Sadhana 27 (2002) 35–58. [12] U. Garain, B.B. Chaudhuri, Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multifactorial analysis, IEEE Trans. Systems Man Cyber. Part C-32 (2002) 449–459.
12. U. Pal, B.B. Chaudhuri Indian script character recognition: a survey Pattern Recognition 37 (2004) 1887– 1899.
13. U. Pal and B.B. Chaudhuri On the development of an optical character recognition (OCR) system for printed Bangla script, Ph.D. Thesis, 1997.
14. U. Pal and BB choudhury, Printed Devnagari script OCR system, Vivek 10 (1997) 12–24.
15. Galileo reading system, http://www.sensorytools.com/ galileo.html.

About Authors

Sh. Omar Khan Durrani, Second year MTech student, Deptt. Of Computer Engineering, NITK., Surathkal, Karnataka.
E-mail : pcs03879@nitk.ac.in

Dr. K C Shet is a Professor in Department of Computer Engineering, NITK, Surathkal, Karnataka.
E-mail : kcshet@nitk.ac.in