Vol 10, Issue 1, Jan/2019
ISSN NO:0377-9254

English Transcription of Sanskrit Characters using Predefined Templates

Anand.M, Dr.Sneha Joshi, S.Siva Kumar,
Dept. of H&S
MRIET
Telangana, India

***Abstract*— Most of the languages used in the world today for communication have evolved through spoken form first, and then into written form. There is a great repository of scientific as well as spiritual knowledge present in the form of Sanskrit literature, which could be very much useful in major fields of Computer Science. However, Sanskrit literature is found written in advanced script known as Devanagari, which makes use of both phonetics and special diacritical marks. In this paper, we emphasize on applying image processing practices to the input image contacting printed Sanskrit characters, in order to recognize the basic blocks of Sanskrit language. The proposed work deals with transcription of Sanskrit characters written in Devanagari script into English script using user-defined templates for each character and mapping a character from an image input to its appropriate transcript form.**

Keywords*— *Sanskrit characters, Devanagari, Transcription, Transliteration.

I.INTRODUCTION

Knowledge needs written form for its representation. Language serves as a carrier for passing on the knowledge from one generation to the next generation. All of the languages originated in oral form first, and then evolved in written form. There are various ancient languages that stood the test of time and passed the knowledge generation after generation. One of the most ancient languages known to human being is Sanskrit. The oldest known source of knowledge, Vedas, is written in Sanskrit, which contain nearly every knowledge system required by humans to live. These systems, for e.g. Ayurveda, have evolved as per the need of time and ability of common people to understand. Various Indian Sages have played important role in transforming this knowledge into simple forms through their literature. Be it any field today, the richness, power, accuracy, simple and unambiguous grammar and the amount of knowledge available in Sanskrit makes it an ideal candidate to be combined with Computers.

Sanskrit is based on ancient script known as Devanagari. It uses “diacritical” marks, the little marks which are present above or below certain characters, which impart certain sound & meaning to them. In order to learn a new language, first its script is to be learnt. Normally, human beings learn to speak new language by first understanding how the characters of that new language are to be uttered in their mother tongue.

The Sanskrit completely based on sound. Whatever generic and permitted sounds a human need to make for creating utterances; those are clearly defined in Sanskrit. Sound has such importance in Sanskrit that a wrong pronunciation leads to completely different meaning. However, it‟s not a limitation but an advantage as there can be only one meaning for written form, and to convey some meaning there is only one written form. This makes Sanskrit language unambiguous. The Sanskrit character analysis relies on the study of sounds made by specific regions in mouth. These are shown in figure 1.

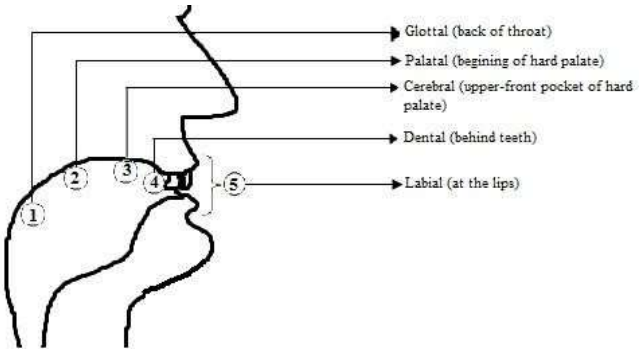

Fig. 1. Anatomy of Phonetics

These well-defined regions are called as Throat (Glottal), Palate (Palatal), Cerebrum (Cerebral), Teeth (Dental) and Lips (Labial). Vowels are the „Voice Patterns" originating at these places, while Consonants are the „form" of these voice patterns. A Vowel can be pronounced alone but a consonant cannot be pronounced without added vowel for

अ = a (vowel) while क = ka (consonant) i.e. □□ + अ.

As there is one-to-one relation between spoken and written form, the context and intention becomes easily clear. Hence correct utterance is a key for correct knowledge deriving from Sanskrit literature. Here, a novel method is described to transcribe the printed Sanskrit text present in the input image into appropriate English script.

II. LITERATURE REVIEW

Considering the efficient features and vast application scope of Sanskrit in major domain such as AI, Knowledge-based Systems and also in NLP studies, a serious thought is to be given to its utilization. There"s a need of some compiler/ interpreter to process vast volumes of Sanskrit literature, which can prove to be very fruitful in advanced technological research. The work carried out here by can prove to be a direction for future work in NLP using Sanskrit. Sanskrit language has a well-defined grammar written by Maharshi Panini [7]. This grammar makes it the most suitable natural language for AI. Due to its unambiguous and detailed grammar, Sanskrit is considered the best natural language to be used. On the other hand, English language poses quite a few problems when used as natural language processing. These problems can be avoided by using Sanskrit language grammar. The method of transcription presented here can prove very useful for future implementation for finding meaning of Sanskrit words and sentences.

The Indian languages are not as easy as English language when it comes to character recognition. This is due to complex structure of these languages. As most of the office work tends to rely heavily upon digital document processing and there is rich repository of knowledge in ancient literature, it is must to consider Devanagari, an ancient Indian script, for character recognition [5]. This paper helps language researchers in the field of Devanagari Optical Character Recognition (DOCR). It also puts light on various techniques available for it. However, the approach for character recognition should be more integrated [2]. The field of character recognition has huge scope due to its applications. Here, it is carried out with the help of input given as image of a character. Devanagari character stored in an image is passed as input to the proposed system. The further processing and identification is performed using neural network implementing feature extraction and post-processing in MATLAB. The approach presented here is novel and highly feasible for practical implementation [4].

Character recognition has worked really well for English language, thereby giving hope for its implementation to produce various applications. However, the same cannot be said yet when it comes to character recognition of Indian languages, due to its complex structure and computation. Due to its vast appeal for digital document processing, Devanagari languages ought to be studied and recognized using currently available techniques.

This paper iterates several techniques which can help researchers working in the field of OCR for recognizing Devanagari character [5]. The handwritten-input in some natural language can prove a mean of communication between human and computer. With the help of available technologies, it should be possible to communicate with computers in Indian languages i.e. Devanagari script. Here, the paper presents an algorithm to identify the input in the form of handwritten Devanagari character by implementing several techniques such as segmentation and neural network. It also accepts character drawn with the help of mouse. This method can be very useful for practical realization of language recognizer in future [3].

Due to its vast usage by Indian people, Devanagari script needs attention of researchers. OCR is a very imp tool for recognizing Indian languages due to their wide usage in banks, offices, organizations. This paper proposes implementation of OCR for Devanagari script recognition with high recognition rate [6]. As Devanagari is the widely used script in Indian languages, there's a need to find a way to make computers read Devanagari text. Here, the effort is made to recognize Devanagari characters using ANFIS, a readymade ANN based tool that is trained first. ANN is trained for feature extraction and matching of characters. Several other techniques such as Histogram, Affine Moment Invariant are also used to get recognition rate up to 98%. The above approach, though complex, can be used for future implementation.

Machine Translation systems have been developed or in developing stage using Sanskrit as source or target language [8]. There are several Sanskrit parsers available like Desika [9], which is general purpose Sanskrit parser which can identify the compound and combined word forms using grammar rules of Panini"s *'Ashtadhyayi'* with database of *'Amarakosa'* and processing from *'Nyaya'* and *'Mimamsa Sastras'* [10]. There is Sanskrit wordnet, which is more than conventional Sanskrit dictionary. Dependency parser for Sanskrit Language uses deterministic finite automata (DFA) for morphological analysis [11].

Recognition of handwritten characters is a major challenge nowadays. AI, Expert Systems have helped to some extent for solving this problem. Devanagari character recognition is even more a difficult task. In this paper it is tried to be solved with the help of multistage feature extraction and classification. It implements several techniques such as

Radon and Euclidean distance transforms and ANN.

III. PROPOSED METHODOLOGY

The following process discusses the approach implemented for recognizing Sanskrit character utterance. Input to the process is provided in the form of an image containing printed Sanskrit characters. It can also be given in the form of audio input or an image containing handwritten character. The ultimate output for any of the above input is always generated as text.

The input image usually contains characters printed in RGB form. Thus, this image is first converted to greyscale. Later, this greyscale image is converted to B/W image using threshold technique. The B/W image contains objects containing black as well as white pixels. Presence of consistent black pixels in an image represents an object. Hence, objects with lesser concentration of black pixels are to be omitted here.

The next step is to compare the objects in the input image with already defined template. This template contains pattern of black pixels for each character along with its utterance. It is created before extraction of characters from input image. Then, black pixels are counted from an input image.

For each character in the image:

- Extract the character
- Convert the image to text with correlation function.

Finally we write the utterance for a character in the form of English text into a text file.

Template Creation

The main focus of our process is to define English utterances of Sanskrit characters to make it possible to read them. For this, it is required to have standard form to define structure and sound of each Sanskrit character. Template for each character is created and stored. It contains specific pattern of black pixels (labelled as 1=black and 0=white) and an utterance associated with it. This template is required in character extraction and utterance generation.

The template is created in such a manner that the bitmap for each character is a bounded region of 42 X 30 pixels. This imparts uniformity to the template and enables better mapping in character extraction step. Faster Mapping enables early utterance detection.

Bitmap for each character along with its utterance is stored continuously in a matrix of size 42 X (30X60) pixels. This is the final step in template creation. It creates a final template file which is then available for further processing.

Character Extraction

The input black-n-white image typically contains multiple characters present in a single line. Each character is supposed to be separated by white space. In this step, these characters are read from single line at a time and then stored in an array.

However, the extracted characters are not necessarily of the same size. It is important for Mapping and utterance generation step that a character appears in fixed, same dimensions. For this, the input image is resized to the size 42 X 30 pixels per character, as in the template file. These resized characters are stored in a separate array.

Character to Text Conversion

Each extracted and resized character, stored in an array, is mapped against the characters present in the template file. The mapping process consists of comparing the bitmap of read characters to the one present in the template. After mapping, character with its maximally associated bitmap along with its utterance is obtained from the template. The utterance for obtained character is written in a text file. The above process is repeated for each and every character found in the input image.

IV. EXPERIMENTAL RESULTS & DISCUSSION

Results

The process starts by accepting input in the form of an image containing more than one character in printed form. The input image is a RGB image. In the next step, this image is converted to greyscale image, which is converted to B/W image. The purpose of this conversion is to increase intensity. To find the information about objects present in an image, intensity information is required. Converting RGB image to greyscale and then B/W improves the intensity of objects in that image, thereby making it easier to spot those objects.

The obtained B/W image contains pixels of only two colors: black and white. The presence and pattern of each character in an image is observed by counting and labeling these pixels. This happens in the next subsequent steps. The characters are extracted from the B/W image and are stored.

Meanwhile, a template is generated with the help of the input image for identify the pattern of the characters, creating bitmap for each character and storing it in a template file with an utterance associated with each character. The utterance is an English pronunciation of that character as per the sound it makes when uttered in Sanskrit.

EXPERIMENTAL RESULTS FOR SANSKRIT CONSONANTS

Input Image	Output Text File
क ख ग घ ङ	KA KHA GA GHA NGA
च छ ज झ ञ	CA CHA JA JHA NYA
प फ ब भ म	PA PHA BA GA MA
श ष स ह ळ	SHA PA SA HA LAA
त थ द घ न	TA THA DA DHA NA
ट ठ ड ढ ण	TTA TTHA DDA DDHA NNA
य र ल व	YA RA LA VA

Fig. 2. Experimental Results for Consonants

The extracted characters, stored in an array, are then mapped in the template file on at a time by comparing the bitmaps of both the characters. The character from a template file, for which there is maximum similarity between the two bitmaps, is then selected as a result. The output is then given as the utterance associated with the selected character from the template file.

The figures 2, 3 and 4 show the experimental results. Each figure represents the result in a table containing two columns. The first column contains the characters extracted from input image, while the second column shows the utterances for each of the characters from first column. Figure 2 shows result of passing image containing Sanskrit consonants as input. There are total 34 consonants which have been passed as a batch of six 5-consonants-a-line images and one 4-consonants-a-line image. Out of 34 consonants, 32 have been identified correctly.

Figure 3 shows result of passing image containing Sanskrit vowels as input. There are total 17 vowels which have been passed as batch of three 5-vowels-a-line images and one 2-vowels-a-line image. Out of 17 vowels, 14 have been identified correctly. Figure 4 shows result of passing image containing Sanskrit numerals as input. There are total 10 numerals which have been passed as batch of two 5-numerals-a-line images. Out of 10 numerals, all have been identified correctly.

EXPERIMENTAL RESULTS FOR SANSKRIT VOWELS

Input Image	Output Text File
अ आ इ ई उ	A AA I II U
ऊ ऋ ॠ ऌ ॡ	UOO RU ROO LRU LROO
ए ऐ ओ औ अं	AE AE AA O AM
अँ अः	AM A

Fig. 3. Experimental Results for Consonants

EXPERIMENTAL RESULTS FOR SANSKRIT NUMBERS

Input Image	Output Text File
० १ २ ३ ४	SHOONYA EKA DVI TRI CHATUR
५ ६ ७ ८ ९	PANCHA SHUT SAPTA ASHTA NAVA

Fig. 4. Experimental Results for Consonants

Discussions

The process explained above, when provided with input image containing printed Sanskrit characters, gives the output shown in figures 2,3 and 4. The performance measure for this process is defined as the accuracy with which the system recognizes a character and outputs the English utterance for it.

The process generated output with 94.117% accuracy when provided with input as a consonant. After giving input as a vowel, the output was generated with 76.470% accuracy. In case of numeric input, the process gave output with 100% accuracy. However, consonants and vowels can also be recognized with even more accuracy, if the system is trained rigorously.

V. CONCLUSION

The proposed approach helps us to identify the basic blocks of Sanskrit language i.e. consonants, vowels and numerals. The method proposed in this paper tries to answer the question „How to pronounce Sanskrit character?" It does so through transcription of each Sanskrit character into English script text which speaks more about its utterance. The process works with numerals accurately while there is still scope in case of consonants and vowels. It can prove to be very much useful in transliteration process while working with

Natural Language Processing applications targeted for Sanskrit language. In future, we wish to incorporate the Neural Network to improve the accuracy.

REFERENCES

[1]. C. Bhole, "Devanagari Character Recognition Using Multistage Classification & Feature Extraction Technique," International Journal of Trend in Research and Development, vol. 3, pp. 164-168, 2016.

[2]. V. Dongre and V. Mankar, "A review of research on Devnagari character recognition," International Journal of Computer Applications, vol. 2, no. 2, pp. 8-15, 2010.

[3]. P. Goyal, S. Diwakar and A. Agrawal, "Devanagari character recognition towards natural human-computer interaction," Proceedings of India HCI, 2010.

[4]. A. Karia, S. Sharma, R. Rodrigues and M. Save, "Character Recognition (Devanagari Script)," International Journal of Engineering Research and Applications, vol. 5, no. 4, pp. 109-113, 2015.

[5]. N. Pratap and S. Arya, "A Review of Devnagari Character Recognition from Past to Future," International Journal of Computer Science and Telecommunications, vol. 3, no. 6, pp. 77-82, 2012.

[6]. R. Singh, C. Yadav, P. Verma and V. Yadav "Optical character recognition (OCR) for printed devnagari script using artificial neural network," International Journal of Computer Science & Communication, vol. 1, no. 1, pp. 91-95, 2010.

[7]. D. Teja and S. Kothuru, "Sanskrit in Natural Language Processing," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 3, pp. 596-600, 2015.

[8]. Raulji J.K., Saini J.R.: Sanskrit Machine Translation Systems: A Comparative Analysis. International Journal of Computer Applications, vol. 136 No. 1, pp. 0975—8887 (2016).

[9]. Desika (Natural Language Understanding System), http://tdil.mit.gov.in/download/Desika.htm.

[10]. "Sanskrit Wordnet" Available on http://www.cfilt.iitb.ac.in/wordnet/webswn/english_version.php

[11]. Antony, P.J.: Machine Translation Approaches and Survey for Indian Languages. Computational Linguistics and Chinese Language Processing, vol. 18, pp. 47-78 (2013).