Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 1; January-March, 2016, pp. 5-8
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Identification of Character Pattern in Devanagari Words for Enhancement of Recognition Accuracy

Manoj Kumar Gupta[1], C. Vasantha Lakshmi[2] and C. Patvardhan[3]

[1,2]Department of Physics and Computer Science Dayalbagh Educational Institute, Dayalbagh, Agra-282005, India
[3]Department of Electrical Engineering Dayalbagh Educational Institute, Dayalbagh, Agra-282005, India
E-mail: [1]mkg0710@rediffmail.com, [2]cvasanthalakshmi@gmail.com, [3]cpatvardhan@gmail.com

Abstract—*From the OCR point of view, the complexity inherent in middle zone Devanagari characters can be managed if the character set of all frequently used connected symbols of overall Devanagari Text is identified. The complexity is further reduces when this frequently used character set is further classified on the basis of structural properties which are invariant across fonts and sizes. Hence the complexity in recognition of unknown character symbol is reduced to a much smaller number of possibilities resulting in enhancement of recognition accuracy.*
Recognition accuracies can further be enhanced by identifying the possible character set for single letter words. The result of the study shows that there are only 17 characters in the middle zone which are occurring in single letter words. Among these 17 characters, more than 50% single letter words are covered by four characters क ह थ म *. An analysis of the coverage of the character set of these 17 characters in the middle zone done on 2462 words shows the 100% coverage by these characters in single letter words.*

1. INTRODUCTION

From the OCR point of view, the recognition accuracies can be enhanced, once the complexities of Devanagari script are identified and reduced by separating the complexities to different manageable chunks.

Devanagari script has core characters in the middle strip and optional modifiers above and/or below core characters. So primarily at first level, the complexity is reduced by separation of core characters in the middle zone and modifiers using various segmentation techniques.

Next level of complexity is in to identify the possible list of characters in the middle zone. Theoretically there can be 46656 conjunct characters formed by combining 36 fundamental characters. A study on 469580 words from a variety of sources shows that there are only 345 symbols in the middle zone which are used more frequently [1]. This reduces complexity limiting to the character set of 345 symbols.

Another complexity in Devanagari is that, the character shape changes drastically with fonts. Hence a classification scheme needs to take care of this complexity. This complexity can be managed by adopting the classification scheme based on the structural properties. This classification scheme reduces 345 frequently used characters into 16 small manageable classes by identifying the various easily distinguishable features, e.g. the presence or absence of the vertical bar and number of places touching to shirorekha [1]. It enhances the recognition accuracy by reducing the errors introduced due to font and size variations [1].

There are large numbers of components or symbols in each word. Due to this, while component level recognizers perform well but the word level and in turn document level accuracies are not acceptable in practical situations [2].

Performance of a recognition system increases if it can detect and correct errors. It may also be necessary to use contextual information. The application of context makes it possible to detect errors and even to correct them. In post processing phase of Devanagari OCR, accuracy is enhanced by using the word dictionary and other suffix/prefix based error correction techniques [3].

Motivation for the present work: The complexity can be further reduced if occurance of more character pattern is identified in the pre processing phase itself. Whether the recognition accuracy can be further increase by identifying the character set which can be present in the single letter word? This is the pertinent question behind the motivation for undertaking this research work.

The paper is organized as follows: Section 2 describes the proposed approach. Section 3 describes the frequency analysis of single letter words. Section 4 describes the coverage analysis of single letter words. Finally, Conclusions are given in Section 5.

2. PROPOSED APPROACH

Typically in an OCR a page of text is read and converted into binaries. The lines are extracted from the text. Then for each line, words are extracted. And then middle zone connected components are extracted from each word. Once the symbols in the middle zone are separated out and features identified for their recognition, the OCR problem is considerably simplified.

There are various types of words of varying length in any Devanagari text where in the frequently used character are 345 but the possible character set for a single letter word can be small. In a single letter word, the number of symbol is either 1 or number of symbol is 2 but another symbol is Bar. The example of single letter words is shown below in Table 1.

Table 1: Example of Single Letter Word

Devanagari Text	Single Letter Word	Character in middle zone of single letter word other than bar
सभी जानवर सोचने लगे कि यह तो रंगा सियार है	कि तो है	क त ह

The proposed approach is to observe the pattern of Devanagari characters occurring in middle zone of the single letter words of the various documents so that character set can be defined for the single letter words.

Identification of character set for single letter words will enhance the recognition accuracy.

Various steps to identify the character in the middle zone of the single letter word are given below in Fig 1.

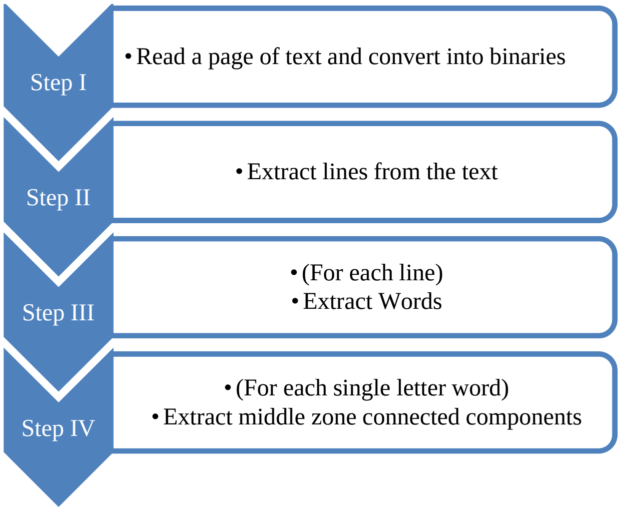
Step I
• Read a page of text and convert into binaries
Step II
• Extract lines from the text
Step III
• (For each line)
• Extract Words
Step IV
• (For each single letter word)
• Extract middle zone connected components

Fig. 2: Steps to identify the characters in the middle zone of the single letter word

3. FREQUENCY ANALYSIS OF SINGLE LETTER WORD

Frequency analysis is done manually to find out the symbol occurring in the middle zone of the single letter word. Frequency analysis is done in four articles of different contexts with different sizes from different sources [4][5][6][7]. The results are summarized in Tables 2, 3, 4 and 5.

Table 2: Results of Frequency Analysis of Single Letter Word in First Article

Sno	Char in middle zone	Frequency of single letter word	%age of single letter word (Total words = 1561)
1	क	115	7.37
2	ह	72	4.61
3	थ	35	2.24
4	म	34	2.18
5	स	24	1.54
6	न	23	1.47
7	त	22	1.41
8	भ	20	1.28
9	व	8	0.51
10	द	6	0.39
11	ज	5	0.32
12	ल	5	0.32
13	अ	4	0.26
14	ख	1	0.06
15	ग	1	0.06
	Total	**375**	**24.02**

Table 3: Results of Frequency Analysis of Single Letter Word in Second Article

Sno	Char in middle zone	Frequency of single letter word	%age of single letter word (Total words = 468)
1	क	26	5.56
2	ह	8	1.71
3	थ	5	1.06
4	म	14	2.99
5	स	9	1.92
6	न	18	3.85
7	त	2	0.43
8	भ	2	0.43
9	व	1	0.21
10	द	5	1.07
11	ल	7	1.50
12	ब	1	0.21
	Total	**98**	**20.94**

Table 4: Results of Frequency Analysis of Single Letter Word in Third Article

Sno	Char in middle zone	Frequency of single letter word	%age of single letter word (Total word = 878)
1	क	67	7.63
2	ह	25	2.85
3	थ	17	1.94
4	म	24	2.73
5	स	24	2.73
6	न	17	1.94
7	त	7	0.80
8	भ	3	0.34
9	व	4	0.45
10	द	3	0.34
11	ज	2	0.23
12	ल	4	0.46
13	ख	1	0.11
	Total	**198**	**22.55**

Table 5: Results of Frequency Analysis of Single Letter Word in Fourth Article

Sno	Char in middle zone	Frequency of single letter word	%age of single letter word (Total word = 1339)
1	क	130	9.71
2	ह	125	9.34
3	म	33	2.46
4	स	39	2.91
5	न	4	0.30
6	त	6	0.45
7	भ	28	2.09
8	व	6	0.45
9	ज	2	0.15
10	य	4	0.30
	Total	**377**	**28.16**

The results show that there are only 17 characters in the middle zone of the single letter word. Among these 17 characters, more than 50% single letter words are covered by क ह थ म characters. A summary is shown below in Table 6.

Table 6: Summary of the Symbols and Percentage of each Symbol

SNO	Char	Article I		Article II		Article III		Article IV	
		Frequency of Single Letter Word	%age of each Symbol among Single Letter Word	Frequency of Single Letter Word	%age of each Symbol among Single Letter Word	Frequency of Single Letter Word	%age of each Symbol among Single Letter Word	Frequency of Single Letter Word	%age of each Symbol among Single Letter Word
1	क	115	30.67	26	26.53	67	33.84	130	34.48
2	ह	72	19.20	8	8.16	25	12.63	125	33.16
3	थ	35	9.33	6	5.10	17	8.59	-	-
4	म	34	9.07	14	14.29	24	12.12	33	8.75
5	स	24	6.40	9	9.18	24	12.12	39	10.35
6	न	23	6.13	18	18.38	17	8.59	4	1.06
7	त	22	5.87	2	2.04	7	3.54	6	1.59
8	भ	20	5.33	2	2.04	3	1.52	28	7.47
9	व	8	2.13	1	1.02	4	2.02	6	1.59
10	द	6	1.60	5	5.10	3	1.52	-	-
11	ज	5	1.33	-	-	2	1.01	2	0.53
12	ल	5	1.33	7	7.14	4	2.02	-	-
13	अ	4	1.07	-	-	-	-	-	-
14	ख	1	0.26	-	-	1	0.51	-	-
15	ग	1	0.26	-	-	-	-	-	-
16	ब	1	0.21	1	1.02	-	-	-	-
17	य	4	0.30	-	-	-	-	4	1.06
	Total	**375**	**100%**	**98**	**100%**	**198**	**100%**	**377**	**100%**

4. COVERAGE ANALYSIS OF CHARACTERS PRESENT IN SINGLE LETTER WORD

An analysis of the coverage of the character set of these 17 characters is done on 2462 words [8]. The overall coverage by these identified 17 characters in single letter word is found to be 100% on these 2462 words. The results of the analysis are summarized in Table 7.

Table 7: Analysis of Coverage of Single Letter Word using Identified 17 Character

Sno	**Source Document**	**Pages**	**Words**	**Coverage by identified 17 character in single letter word**
1	Chapter 1	4	735	100%
2	Chapter 2	4	828	100%
3	Chapter 3	5	899	100%
	Total	13	2462	100%

5. CONCLUSIONS

Frequency analysis is done manually in four articles of different contexts with different sizes from different sources to find out the symbol occurring in the middle zone of the single letter word. The total 4246 words are observed and out of these, 1048 single letter word are found.

The result shows that there are only 17 characters which are occurring in the middle zone of the single letter words. These are क ह थ म स न त भ व द ज ल अ ख ग ब .

Among these 17 characters, more than 50% single letter words are covered by four characters viz. क , ह , थ , म .

An analysis of the coverage of the character set of these 17 characters done on 2462 words shows the 100% coverage by these characters in single letter word.

The utility of the proposed approach is to enhance the recognition accuracies by identifying the character set for the single letter word.

REFERENCES

[1] M. K. Gupta, C. Vasantha Lakshmi, M. Hanmandlu, C. Patvardhan "An Exhaustive Font and Size Invariant classification Scheme for OCR of Devanagari Character" in International Journal on Natural Language Computing, Feb.2015, Vol. 4 No. 1 pp 1-21

[2] V. Rasagna, A. Kumar, C. V. Jawahar, R. Manmatha "Robust Recognition of Documents by Fusing Results of Word Clusters" in 10th International Conference on Document Analysis and Recognition, ICDAR Year 2009, pp 566-570

[3] R. Jayadevan, S. R. Kolhe, P. M. Patil, U. Pal "Offline Recognition of Devanagari Script: A Survey" in IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Year: 2011, Volume: 41, No. 6 pp 782-796

[4] Vishnu Sharma, Panchtantra Ki 101 Kahaniya,, 7th ed., Manoj Publications, Delhi, 2010, pp 9-12

[5] Jagdish Singh, Gyan Sagar 3, Gita Publishing House, Delhi, pp 5-8

[6] Vivek Mohan, Tenaliram Ki Kathain, Jhole Main Katora, 1st ed., Raja Pocket Books,Delhi, 2004, pp 3-7

[7] Dr. Om Prakash Ji Maharaj, Ayurved Ka Chamatkar Haldi, Laxmi Prakashan, Delhi, pp 7-10

[8] Jagdish Singh, Gyan Sagar 6, Gita Publishing House, Delhi, pp 5-17