AN EXHAUSTIVE FONT AND SIZE INVARIANT CLASSIFICATION SCHEME FOR OCR OF DEVANAGARI CHARACTERS

Manoj Kumar Gupta[1], C. Vasantha Lakshmi[2], M. Hanmandlu[3], C. Patvardhan[4]

[1]Ph.D. (C.S.) Scholar, Department of Physics and Computer Science, Dayalbagh Educational Institute, Dayalbagh, Agra-282005, India
[2]Department of Physics and Computer Science, Dayalbagh Educational Institute, Dayalbagh, Agra-282005, India
[3]Department of Electrical Engineering, I.I.T. Delhi, New Delhi-110016, India
[4]Department of Electrical Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra-282005, India

ABSTRACT

Main challenge in any Optical Character Recognition (OCR) system is to deal with multiple fonts and sizes. In OCR of Indian languages, one also has to deal with a huge number of conjunct characters whose shape changes drastically with fonts. Separating the conjunct characters into its constituent symbols leads to segmentation errors. The proposed approach handles both the above listed problems in the context of Devanagari script. An attempt is made to identify all possible connected symbols of Devanagari (could be a consonant, vowel, half consonant or conjunct consonant henceforward shall be referred as a basic symbol) in the middle zone without segmenting the conjunct characters. On observing 469580 words from a variety of sources in our study, it is found that only 345 symbols are used more frequently in the middle zone and cover 99.97% of the text. They are then classified into 16 different classes on the basis of structural properties which are invariant across fonts and sizes. To validate the proposed classification scheme, results are presented on 25 fonts and three sizes.

KEYWORDS

Optical Character Recognition, Conjunct Character, Structural Features, Header line (Shirorekha), Vertical Bar

1.INTRODUCTION

In a typical OCR system, the input document is scanned by an optical scanner to produce a gray level or binary bit mapped image. It is a common practice in OCR applications to convert image into binary for easy processing. The image is then segmented into lines, words and isolated characters.

After segmentation, characters in the form of binary matrices are subjected to smoothing, noise elimination, size normalization, and maintenance of the aspect ratio to facilitate the extraction of features in the subsequent stage. Identification of characters is achieved by comparing the extracted features with the features obtained from the set of samples used in the learning phase. A typical block diagram is shown as Figure 1

DOI : 10.5121/ijnlc.2015.4101

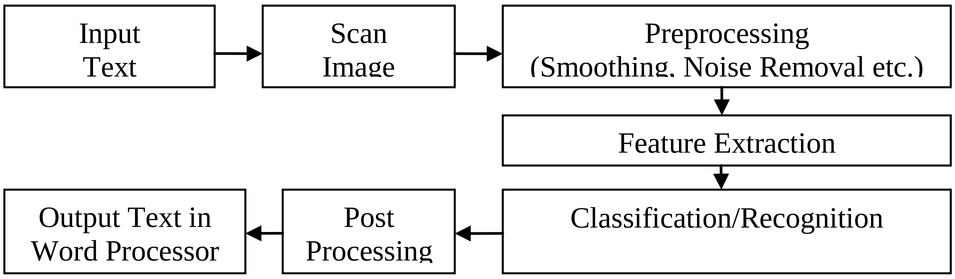


Figure 1. Block Diagram of a standard feature based OCR

The main preprocessing techniques are smoothing, normalization and segmentation Smoothing performs the functions of filling and thinning. Filling eliminates small breaks, gaps and holes. Thinning is the process of reducing the width of a line comprising many pixels to just one pixel. Normalization involves an algorithm that not only corrects the slant of the individual character or the whole word but also modifies the size of a character size to the standard size.

In the Feature Extraction stage, features that can discriminate one class of characters from others are considered. These can be pixel-based, structure-based or stroke-based.

Post processing is usually done to increase the performance of a recognition system. The performance 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 and correct errors. OCR has been known to be a difficult problem although making a computer read documents just as the humans do it has been a dream for long. Some of the difficulties in creating "perfect" OCR systems (or at least systems with human level performance) are as follows.

a. Documents can be created with multiple fonts and sizes and with different styles like bold, italics, strikethrough etc. Some of these may appear to be quite different from the others.
b. Text may be written at different orientations and even or curved lines.
c. Recognition of text and data in tabular format provides its own difficulty.
d. Noise in the document images may corrupt the character shapes and make them difficult to recognize.
e. Scanning errors may result in characters that touch each other making segmentation difficult.

The above problems have resulted in making even development of OCR systems for relatively simple Roman Script a difficult problem. Commercial software performs well on "well behaved' or "OCR friendly" documents but not otherwise.

Devanagari script has 13 vowels and 36 consonants resulting in a total of 49 basic characters. Apart from these, there are compound characters which are formed by combining two or more basic characters. The shape of compound character is usually more complex than the constituent basic characters. The individual characters are joined by a header line called "Shiro Rekha". Various types of Devanagari characters are depicted in Table 1.

Table 1. Examples of Various Types of Devanagari Characters

Sno	Type	Example
1	Vowels	अ आ इ ई उ ऊ ऋ ए ऐ ओ औ अं अः
2	Modifier symbols corresponding to the vowels	ा ि ी ु ू ृ े ै ो ौ
3	Consonants Modifier symbols attached with consonant क	क ख ग घ ङ च छ ज झ ञ ट ठ ड ढ ण त थ द ध न प फ ब भ म य र ल व श ष स ह क्ष त्र ज्ञ का कि की कु कू कृ के कै को कौ
4	Pure Consonants	क् ख् ग् घ् ज् झ् ञ् ण् त् थ् ध् न् प् फ् ब् भ् म् य् ल् व् श् ष्
5	bi-consonantal conjuncts	क्च क्न ख्व च्च ज्ब थ्स त्क न्त
6	tri-consonantal conjuncts	क्स्ट ञ्ज्य त्स्य न्ध्य ध्ल्य म्प्य
7	Conjunct with modifier 'ra'	फ्र ट्र ज्र ष्क्र न्द्र म्ब्र
8	Conjuncts with ह	ह्य ह्म ह्व ह्ल ह्न ह्र

The script has core characters in the middle strip comprising of basic characters or compound characters along with optional vowel modifiers above and/or below core characters. They can be divided into 3 zones and a header line as shown in Figure 2

Figure 2. Example of Various Zones in Devanagari Character

Motivation for the present work: In OCR of Indian language, a fundamental issue that needs to be addressed is: what is the basic unit of recognition? Whether one should attempt to split the conjunct character into its constituent basic characters or conjunct character should be treated as a whole. Also whether one should deal with all possible characters or only characters that are used generally. Is it worthwhile to go for the first level of classification using the structural properties that do not vary across fonts and sizes? These are pertinent questions behind the motivation for undertaking this research work.

The paper is organized as follows: Section 2 describes the challenges in Devanagari OCR and representative approaches. Section 3 presents the proposed approach. Identification of basic symbols is dealt in Section 4. The results of frequency and coverage analysis of basic symbols are summarized in Section 5 of this paper. Section 6 develops a classification scheme. Validation of the proposed classification scheme over 25 fonts is discussed in Section 7. Finally, Conclusions are given in Section 8.

2. Challenges in Devanagari OCR And Representative Approaches

All the difficulties highlighted in Section 1 on OCR in general hold true even for Indic Scripts that pose more challenges due to the unique writing styles permitted in these scripts.
The complexity in Devanagari script arises due to the presence of a large number of symbols. Some of them have similar shapes leading to misclassifications. In particular, the main sources of difficulty pertaining to Devanagari script, apart from those listed above as generic causes are as follows.

a. The shapes of the characters may vary drastically across fonts especially in the case of conjunct characters.
b. The symbols employed in the script are not restricted to the basic vowels and consonants but also to the enormous number of conjunct characters generated by the combination of the simple vowels and consonants with different marks and half characters estimated to run into several hundred thousand.

Various methods used in the past for OCR of Devanagari text include the use of shape based features, separation of pattern classes into several groups based on some similarity of characters, Statistical features, and gradient and curvature based feature extraction methods [1].

Kompalli et al. [2] use the gradient features to classify segmented images into 74 classes: 4 ascenders, 2 descenders, and 68 core components. The 68 core components contain 36 vowels and consonants and 32 frequently occurring conjuncts. A nearest neighbour classifier is used for ascenders and descenders. Core components are pre-classified into four groups based on the presence of a vertical bar, no vertical bar, vertical bar at the centre, vertical bar at the right or at multiple locations

In the rule-based segmentation approach [3], the shirorekha is determined using projection profile and run length. Structures above the shirorekha are isolated as ascenders. Average character height is used to predict a region of interest (ROI) to locate descender. The ROI is examined using run-length analysis to separate descender from core components.

The recognition-driven segmentation by Kompalli et al. in [3] moots the recognition hypothesis which involves the use of non-linear boundaries to separate Devanagari components, and classifier hypothesis for frequently occurring characters (consonants/vowels or consonant-ascenders) before attempting at the segmentation of conjuncts and consonant-descenders on the basis of the classifier results.

Chaudhuri and Pal [4] assign a character to one of the three groups namely basic, modifier and compound character group on the basis of bounding box width, number of border pixels per unit width and accumulated curvature over the border per unit. Feature based approach is used for basic and modifier character recognition while a combination of feature based template matching approach is used for the compound character recognition.

Bansal and Sinha [5] divide the character set of Devanagari script into three groups based on the coverage of the region of the core strip viz. FULL BOX e.g. अ , UPPER HALF BOX e.g.  and LOWER HALF BOX characters e.g. . A total 62 symbols is identified by them in these three groups. The conjuncts are segmented into character and half character and are grouped into FULL BOX, UPPER HALF BOX and LOWER HALF BOX characters. The FULL BOX characters are further divided into three groups based on the presence and position of vertical bars, namely end bar characters, middle bar characters and no bar characters. End bar characters are further subdivided into two groups based on number of junctions with header line.

Hanmandlu et al. [6] have described a system to classify off-line handwritten Hindi characters into several groups based on some similarity measure. They have identified the header line based on end points and pixels positions in the top half part of the character image. After the identification and removal of header line, all the characters are coarsely classified according to the space between the components of the same character (connected and non-connected characters), presence and position of the vertical bar (End, Mid and No Bar) and presence of closed loops.

3. THE PROPOSED APPROACH

The conjuncts are formed by adding consonant and vowel marks to the consonants. There is a variety of ways in which this is done. Some modifiers or their portions could be above the header line, some added to the character below the character and some in the middle zone. The position and the way in which the marks are joined to the base character may vary from font to font. In some fonts the vowel or consonant modifier may touch the character it modifies and in others it may not. All these differences create complications in segmentation. What works for one font may not be so for another.

Dividing the words into upper, middle and lower zones reduces the count of characters present in the middle zone and which appear to be connected. Theoretically there can be 1296 (36x36) biconsonantal conjunct characters formed by collating 36 fundamental symbols. Further there can be 46656 (1296x36) triconsonantal conjunct characters formed by combining the 1296 biconsonantal conjuncts with 36 fundamental symbols. This indicates the possibility of presence of fairly a large number of connected symbols in the middle zone.

In the development of OCR, a fundamental issue that needs to be resolved is level of recognition. We can go for the recognition of the connected components that are left after the removal of the Shirorekha or smaller sub-units with the middle line segment, ascenders and descender separately or the constituent basic characters of a conjunct character or the conjunct character as a whole?

It is really difficult to separate a conjunct into a consonant and pure consonant as it is not always that the two are connected exactly at the middle of the conjunct. Problems in Ascender segmentation occur when ascenders get connected with each other. The position of descender changes with alphabet height and font. In the recognition-driven segmentation paradigm Devanagari characters originate from fusion left to right and top to bottom and are to be segmented along nonlinear boundaries. The modifiers are separated from the character and recognized separately in Indian languages. In this, segmentation takes place without an understanding of what is to be expected as the outcome of recognition when it is not performed but only the segmentation is performed for recognition.

Owing to the presence of default touching characters in Devanagari, segmentation of conjuncts is inefficient and highly font dependent as the position of touching is not fixed in any two conjuncts.

The total number of conjuncts to be segmented and recognized being unmanageable from OCR point of view it is necessary to devise ways and means of bringing their number down to manageable levels.

Primarily the complexity lies in the middle zone. One way to do this is to consider only the conjuncts occurring in the middle zone i.e. removing all the symbols including and above the header line ("Shirorekha") and also the modifiers attached at the bottom. This relevant portion is clearly shown in Figure 2.

Vasantha Lakshmi and Patvardhan [7] have dealt with the basic symbols for Printed text (TOSP) in Telugu OCR System. A Basic Symbol is defined as a single connected entity in Telugu script and is treated as the unit of segmentation and recognition.
A similar approach can be adapted for Devanagari with a difference that only the connected components in the middle zone are treated as units of recognition. An attempt is made to identify how many basic connected symbols can be present in the middle zone in addition to finding their distinguishing characteristics such as structure, shape and presence/absence of certain character segments.

Once the symbols in the middle zone are separated out and features identified for their recognition, the OCR problem is considerably simplified because the other portions, i.e. the top zone (above the header line) and the bottom zone can then be separately processed. The number of symbols to be handled at each stage remains within manageable levels. This approach therefore offers a way to mitigate the complexities of handling a huge number of conjunct characters.

An analysis is done to identify the basic symbols that are used more often. Structural properties of these symbols that are invariant across fonts and sizes are identified and classification is performed using these invariant structural properties as far as possible. This approach makes it much easier to create a classifier.

This approach brings out certain interesting observations listed here.

a. Contrary to our expectations, the number of unique conjunct characters in the middle zone is not too large.
b. Further a small number of symbols provide a very high percentage of coverage. Therefore, if means can be devised to segment and recognize these symbols accurately, considerable part of the OCR problem could be regarded as solved.
c. The conjunct characters occurring in the middle zone can be classified into small manageable classes by means of identification of some easily distinguishable features, e.g. the presence or absence of the vertical bar and so on. The advantage of using these easily recognizable features is that being gross features they remain the same across fonts and sizes.

Building a classifier capable of recognizing the conjuncts in each class and consisting of a relatively small number of symbols is much less difficult and therefore less error prone.

4. IDENTIFICATION OF BASIC SYMBOLS IN THE MIDDLE ZONE

To generate an exhaustive list of basic symbols in the middle zone, several weekly magazines[8][9], newspapers [10][11], children's' magazine[12][13], Essay books[14], books on ayurvedic tips[15], books on cooking tips[16], vyakaran books[17] and other children's course text books [18][19][20] are probed. A list of variety of sources used for identification of symbols is shown in Table 2.

Table 2. List of Variety of Sources used for Identification of symbols

SNO	Source Document	No. of Pages	No. of Words in a page (appx.)	Total No. of Words (appx.)
1	Weekly Hindi Magazine ‘Sukravar’, 18 dec-24 dec, 2010	78	900	70200
2	Weekly Hindi Magazine, ‘India Today’, 23 march, 2011	64	600	38400
3	Hindi News Paper, ‘Sunday Navbharat Times’, 10 Nov. 2013	22	2500	55000
4	Hindi News Paper, ‘ Rashtriya Sahara’, 14 Nov., 2013	24	2000	48000
5	Panchtantra ki 101 kahaniya	170	300	51000
6	Tenaliram ki kathain	38	170	6460
7	Hindi nibandh avam patra lekhan	256	280	71680
8	Ayurved ka chamatkar haldi	80	250	20000
9	Hawkins Akhil Bhartiya pak pustika	81	400	32400
10	Manak vyakaran vyavahar, part-I	184	270	49680
11	Gyan Sagar 6	70	260	18200
12	Gyan Sagar 3	34	120	4080
13	Satrangi hindi pathya pustak 1	56	80	4480
	Total			469580

It is found from a variety of sources that out of 469580 words there are only 345 symbols as shown in Table 3 in the middle zone used frequently. It also includes symbols with half characters disconnected from the character.

Table 3. List of 345 Devanagari Symbols in the Middle Zone

Type	**No. of char**	**Characters**
Single char	42	अ इ उ ऊ ऋ ए क ख ग घ ङ च छ ज झ ञ ट ठ ढ ण त थ द ध न प फ ब भ म य र ल व श ष स ह क्ष त्र ज्ञ श्र
Starting With क्	21	क्च क्न क्ल क्व क्क क्थ क्ख क्प क्म क्य क्स क्फ क्त क्ट क्श क्त्र क्ज क्ड क्स्ड क्स्ट क्स्थ
Starting With ख्	7	ख्त ख्व ख्म ख्स ख्य ख्श ख्भ
Starting With घ्	3	घ्न घ्व घ्य
Starting With च्	5	च्च च्व च्य च्छ च्क
Starting With ज्	7	ज्ज ज्ब ज्व ज्झ ज्म ज्य ज्ज्व
Starting With ञ्	5	ञ्ज ञ्च ञ्झ ञ्छ ञ्ज्य
Starting With त्	19	त्त त्व त्ज त्व त्म त्स त्प त्य त्र त्य त्स्य त्स्प त्थ त्क्ष त्क त्फ त्श त्ल त्स्न
Starting With थ्	3	थ्व थ्य थ्स
Starting With द्	4	द्ध द्व द्य द्द
Starting With ध्	4	ध्न ध्व ध्य ध्म
Starting With न्	26	न्ज न्त न्न न्व न्त्र न्म न्स न्य न्भ न्थ न्ध न्ग न्ट न्ड न्द न्ह न्क न्त्य न्त्र्य न्ध्य न्त्व न्च न्स्ट न्श न्फ्य न्स्ड
Starting With प्	8	प्त प्ट प्न प्ल प्स प्प प्य प्श
Starting With फ्	10	फ्म फ्य फ्स फ्थ फ्त फ्ल फ्क फ्ट फ्ज फ्फ
Starting With ब्	13	ब्ज ब्त ब्ल ब्ब ब्म ब्स ब्य ब्भ ब्थ ब्ध ब्ट ब्द ब्ल्य
Starting With भ्	2	भ्य भ्व
Starting With म्	19	म्त म्न म्ल म्ब म्म म्स म्य म्प म्ड म्द म्ह म्भ म्थ म्क म्फ म्व म्ज म्प्य म्प्ल
Starting With य्	1	य्य
Conjunct Char with ‘ra’	27	क्र ग्र घ्र ड्र ज्र थ्र द्र ध्र ट्र प्र फ्र ब्र भ्र म्र व्र स्र ष्क्र न्द्र ष्प्र स्क्र स्प्र न्फ्र न्ध्र त्प्र म्ब्र म्फ्र म्प्र
Starting With ल्	22	ल्ज ल्ल ल्ब ल्व ल्म ल्स ल्प ल्य ल्थ ल्भ ल्ट ल्ड ल्द ल्ह ल्ग ल्क ल्फ ल्श ल्म्स ल्च ल्त ल्क्य
Starting With व्	4	व्ल व्व व्य व्ह
Starting With श्	2	श्ध श्च
Starting With ष्	11	ष्व ष्म ष्ण ष्प ष्य ष्क ष्फ ष्ड ष्ट ष्ठ ष्च
Starting With स्	22	स्त स्न स्ल स्च स्ब स्व स्त्र स्म स्स स्ख स्प स्य स्थ स्क स्फ स्ट स्त्व स्त्य स्स्व स्थ्य स्ज स्ड
Starting With क्ष्	2	क्ष्य क्ष्म
Starting With त्र्	1	त्र्य
Starting With ह्	6	ह्य ह्म ह्व ह्ल ह्न ह्र
Other char	8	रू रु ः ळ ऽ ॐ ा [illegible]
Half char	41	क् ख् घ् च् ज् ञ् त् थ् ध् न् प् फ् ब् भ् म् य् ल् व् ष् स् क्ष् त्र् ज्ञ् क्स् ञ्ज् त्व् त्स् त् न्प् न्त् न्त्र् त्ध् न्स् ब्ल् म् ल्म् ल्क् स्स् स्त् स्थ् स्व्

5. FREQUENCY AND COVERAGE ANALYSIS OF BASIC SYMBOLS

Frequency analysis is done manually on the identified symbols to find out the most frequently used symbols in two articles on different contexts with different sizes from a weekly Hindi magazine [21][22]. The results are summarized in Tables 4 and 5.

Table 4. Results Of Frequency Analysis OF First Article

Sno	Char	Frequency	%age	Sno	Char	Frequency	%age
1	ा	1560	29.48	43	ध्र	3	0.06
2	क	388	7.33	44	प्र	18	0.34
3	र	287	5.42	45	्र	21	0.37
4	स	236	4.46	46	द्ध	12	0.23
5	ह	222	4.19	47	क्र	5	0.09
6	म	219	4.14	48	द्र	8	0.15
7	न	210	3.97	49	द्व	3	0.06
8	त	193	3.65	50	रू	2	0.04
9	ल	165	3.12	51	त्त	4	0.08
10	ग	136	2.57	52	फ्र	1	0.02
11	प	135	2.55	53	श्व	8	0.15
12	व	135	2.55	54	स्स	2	0.04
13	अ	116	2.19	55	स्थ	6	0.12
14	य	106	2.00	56	स्त	16	0.30
15	ज	98	1.85	57	स्प	1	0.02
16	ब	98	1.85	58	स्य	3	0.06
17	द	100	1.89	59	क्स	1	0.02
18	इ	93	1.76	60	क्ल	2	0.04
19	ट	76	1.44	61	क्क	2	0.04
20	श	72	1.36	62	ष्ट	3	0.06
21	भ	61	1.15	63	ष्य	3	0.06
22	ड	61	1.15	64	न्न	2	0.04
23	ए	49	0.93	65	न्त	7	0.13
24	ख	40	0.76	66	न्य	1	0.02
25	ध	30	0.57	67	ल्ड	2	0.04
26	ढ	30	0.57	68	ल्ल	1	0.02
27	ण	28	0.53	69	ल्स	2	0.04
28	क्ष	20	0.38	70	ल्प	1	0.02
29	उ	20	0.38	71	ल्क	1	0.02
30	ष	20	0.38	72	त्व	1	0.02
31	च	31	0.56	73	त्प	1	0.02
32	घ	11	0.21	74	त्म	1	0.02
33	थ	12	0.23	75	ज्य	1	0.02
34	ठ	10	0.22	76	फ्त	3	0.06
35	ज्ञ	7	0.13	77	त्य	1	0.02
36	श्र	3	0.06	78	स्ट	4	0.08
37	फ	22	0.42	79	:	1	0.02
38	त्र	20	0.38	80	त्क	1	0.02
39	छ	11	0.21	81	म्म	1	0.02
40	झ	4	0.08	82	म्प्यु	1	0.02
41	ऊ	1	0.02	83	स्र	1	0.02
42	द्य	1	0.02	**Total**		**5291**	

Table 5. Results Of Frequency Analysis Of Second Article

Sno	**Char**	**Frequency**	**%age**	**Sno**	**Char**	**Frequency**	**%age**
1	ा	454	29.02	28	क्ष	3	0.19
2	क	122	7.80	29	उ	2	0.13
3	र	122	7.80	30	च	17	1.09
4	न	80	5.12	31	थ	7	0.45
5	स	65	4.16	32	श्र	7	0.45
6	ह	68	4.35	33	फ	5	0.32
7	म	78	4.99	34	प्र	1	0.06
8	त	47	3.01	35	क्र	7	0.45
9	ल	44	2.81	36	त्म	1	0.06
10	ग	18	1.15	37	स्थ	1	0.06
11	प	35	2.24	38	क्ल	3	0.19
12	व	27	1.72	39	स्व	3	0.19
13	अ	33	2.11	40	स्प	2	0.13
14	य	36	2.30	41	्र	1	0.06
15	ज	31	1.98	42	द्र	3	0.19
16	ब	30	1.92	43	ब्द	2	0.13
17	द	23	1.47	44	स्ट	6	0.38
18	इ	18	1.15	45	फ्र	5	0.32
19	ट	25	1.60	46	रू	4	0.26
20	भ	27	1.73	47	न्य	2	0.13
21	ए	23	1.47	48	ल्ल	2	0.13
22	ड	16	1.02	49	क्स	1	0.06
23	ख	16	1.02	50	स्र	6	0.38
24	श	12	0.77	51	घ्य	1	0.06
25	ध	9	0.58	52	म्म	1	0.06
26	ढ	1	0.06	53	श्ध	1	0.06
27	ण	3	0.19	54	ह्द	1	0.06
				Total		**1564**	

The results show that around 97% text is covered by the single characters and the presence of the conjunct characters in these articles is around 3% only. A summary is shown in Table 6.

Table 6. Summary of the Symbols and Percentage of the Text Covered

Text Covered (approx.)	**No. of Symbols**	**Symbols**
97%	42	ा क र न स ह म त ल ग प व अ य ज ब द इ ट भ ड श ख ए ध ढ ण क्ष उ ष च घ थ ठ ज्ञ श्र फ त्र छ झ ऊ द्य
3%	47	ध्र प्र ्र द्ध क्र द्र द्व रू त्त फ्र थ्व स्स स्थ स्त स्प स्य क्स क्ल क्क ष्ट ष्य न्न न्त न्य ल्ड ल्ल ल्स ल्प ल्क त्व त्प त्म ज्य फ्त त्य स्ट ः त्क म्म म्प्यु स्त्र स्व ब्द स्र घ्य श्ध ह्द

An analysis of the total coverage of the text using these identified 345 symbols is done on 5 sample-documents downloaded from the CIIL website [23]. The results of the analysis are summarised in Table 7. The overall coverage by these identified 345 symbols is found to be 99.97% on 13113 words in 5 sample-documents. Details are shown in Table 7.

Table 7. Analysis of Coverage of Text using Identified 345 Symbols

SNO	Source Document	Pages	Words	Not Covered	No. of words in which Symbol not covered	Coverage by identified 345 symbols
1	Hindi1.pdf	5	713	ठ्य	2	99.72 %
2	Hindi2.pdf	10	2026	-	-	100%
3	Hindi3.pdf	15	4665	-	-	100%
4	Hindi4.pdf	15	2668	-	-	100%
5	Hindi5.pdf	17	3041	ड्ड	2	99.93%
	Total	62	13113			99.97%

6. CLASSIFICATION SCHEME

On the basis of analysis in the previous sections 345 basic symbols are identified and these provide very high coverage of 99.97%. A detailed classification scheme is developed in this section using the structural properties. This classification scheme is required to be robust across fonts and sizes. Once such a classification scheme is created for each of the individual classes, this scheme is expected to be more tolerant to noise and robust across multiple fonts and sizes.

Various categories for BAR Type are identified as follows:

1. Some characters have a vertical bar at the end of them.
2. Some characters have a vertical bar in the middle of them.
3. Some characters do not have any bar.
4. Some conjuncts have two vertical bars with them.
5.

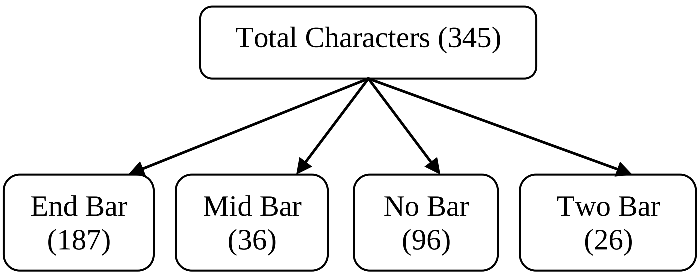


Figure 3. Classification based on presence and location of vertical bar

The way the characters touch the header line is different; a character may touch only once with the header line or more than once. This categorization of the count of touching is as follows:
1. Characters not touching the header line.
2. Characters touching the header line once.
3. Characters touching the header line at two places.
4. Characters touching the header line at three places.
5. Characters touching the header line at four places.
6. Characters touching the header line at five places.

Based on these properties, there are 16 different character classes to which a character may belong.

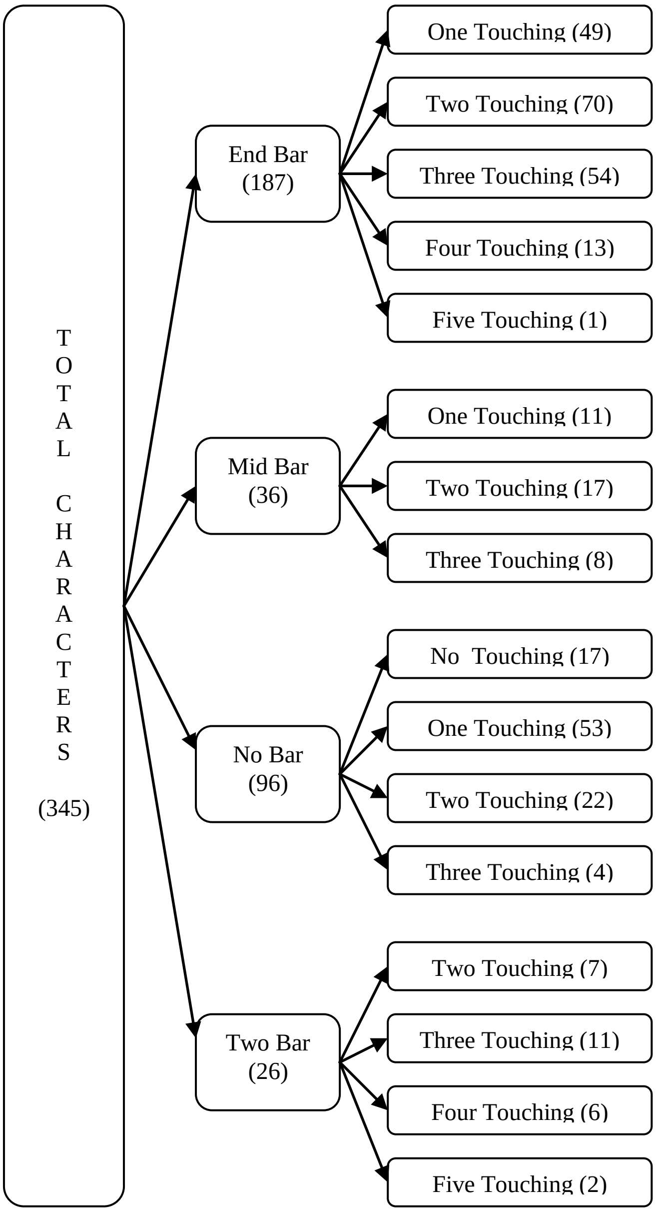


Figure 4. Classification based on touching count

The characters in each of the 16 classes are shown in Table 8.

Table 8. Classification of Total 345 Character based On Bar Type and Touching Count

SNO	Property	No. of char	Characters
1	End Bar One Touching	49	ज ञ त न ल ज्ञा त्त ज्ज ञ्ज त्ज न्ज न्त न्न ब्ज ब्त ब्ल ल्ज ल्ल व्ल ग च ब व त्र त्न ज्ञ ब्र व्र च्च च्व ज्ब ज्व ञ्च त्व न्व न्त्र ब्ब ल्ब ल्व व्व ज्ज्व त्त्व न्त्व ल्च ल्त त्ल न्च
2	End Bar Two Touching	70	अ झ म स स ख्त म्त म्न म्ल स्त स्न स्ल घ्न प्न प्ल य द्य ह्य ह्म म्र ख्ब ख्व म्ब स्च स्ज स्ब स्व स्त्र स्त्व घ प ष घ्र प्र घ्व ष्व ज्झ ज्म ञ्झ त्म त्स न्म न्स ब्म ब्स ल्म ल्स च्य ज्य त्प त्य न्य ब्य ल्प ल्य व्य ञ्य न्त्य न्ञ्य त्न्य स्र ञ्ज्य म्व म्ज त्स्न त्प्र म्ब्र ब्ल्य ष्च ग्र
3	End Bar Three Touching	54	ल्म्स त्क्ष न्भ ब्भ ख्म ख्स म्म म्स स्म स्स श्र भ क्ष ध्न प्स ष्म त्स्य त्स्प त्थ न्थ न्ध ब्थ ब्ध ल्थ ल्भ ख्य म्य म्प य्य स्ख स्प स्य स्त्य स्स्व घ्य प्प प्य ष्प ष्य थ ध थ्व ध्व थ्र ध्र भ्र भ्व न्ध्र श्ध श्च स्प्र म्प्र म्प्ल ष्प्र
4	End Bar Four Touching	13	ख्भ म्भ क्ष्म म्थ स्थ न्ध्य थ्य ध्य भ्य क्ष्य ध्म म्प्य थ्स
5	End Bar Five Touching	1	स्थ्य
6	Mid Bar One Touching	11	क क्र क क् ऋ क्त च्क त्क न्क ल्क ल्क
7	Mid Bar Two Touching	17	फ फ्र फ फ् क्ट त्फ ल्फ म्क स्क ष्क न्फ्र न्फ क्ट क्ड म्फ स्फ्र ष्क्र
8	Mid Bar Three Touching	8	फ्ट क्श म्फ स्फ ष्फ क्स्ड क्स्ट म्फ्र
9	No Bar No Touching	17	च ज ञ त न ब ल व ञ ज़ न्ट न्ञ त्र ञ्ज ब्ल त :
10	No Bar One Touching	53	इ उ ऊ ट ठ ड ढ द र ह ख ट स्ट ढ्ढ द्व ट्र ड्र द्र ह्व ह्ल ह्न ह्न ळ रू रु छ ॐ न्ग न्ट न्ड न्द न्ह ब्ट ब्द ल्ट ल्ड ल्द ल्ह व्ह ल्ग च्छ ञ्छ ऽ ६ ८ ७ ल्ङ त्ट स्ट स्ठ न्ट न्द्र द्ब
11	No Bar Two Touching	22	म्ड म्द म्ह स्ट ए २ ६ क्ष ए २ प्ट ष्ड ए ष्ठ त्२ ल्२ न्स्ट न्२ न्स्ड स्ड स्ट न्६
12	No Bar Three Touching	4	ख२ प२ ष्प स्२
13	Two Bar Two Touching	7	क्च क्न क्ल क्व क्क क्त्र क्ज
14	Two Bar Three Touching	11	फ्त फ्ल फ्क क्ख क्प क्म क्य क्स क्फ फ्ज ल्क्य
15	Two Bar Four Touching	6	फ्म फ्य फ्स क्थ फ्फ न्फ्य
16	Two Bar Five Touching	2	फ्थ क्स्थ

7.VALIDATION OF PROPOSED CLASSIFICATION SCHEME OVER 25 FONTS

To validate the proposed classification scheme, a system is developed in JAVA. The test data of all 345 symbols is created for the 25 fonts shown in Table 9.

Table 9. List of Font Name For The Test Data

Sno	Font Name	Font Size	Sno	Font Name	Font Size
1	Mangal	24px, 16px, 12px	14	Kruti Dev 714	24px
2	GIST-DVOTAkshar	24px, 16px, 12px	15	Richa	24px
3	GIST-DVOTMaya	24px, 16px, 12px	16	Aparajita	24px
4	GIST-DVOTKishore	24px	17	Kokila	24px
5	GIST-DVOTSubodh	24px	18	Kundali	24px
6	GIST-DVOTVineet	24px	19	Arjun	24px
7	AkrutiDevYogini,	24px	20	Kanika	24px
8	Arial Unicode MS	24px	21	Devlys 010	24px
9	Utsaah	24px	22	Devlys 140	24px
10	Gargi	24px	23	AkrutiDevPriya	24px
11	Gurumaa	24px	24	AkrutiDevGangal	24px
12	Kruti Dev 010	24px	25	AkrutidevPriyanka	24px
13	Kruti Dev 040	24px			

The Figure 5 shows the image of one test data file of Mangal font with 24px font size used for testing. Similar files of different fonts and sizes are created and used for testing.

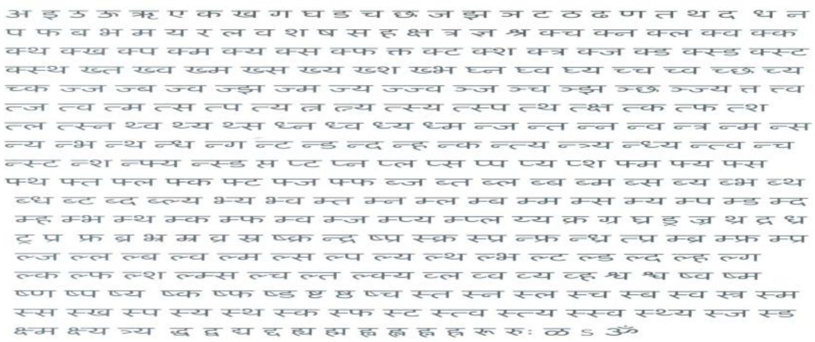


Figure 5. Image of Test Data File of Mangal Font with 24px Font Size

The screen shot of the main menu is shown in Figure 6.

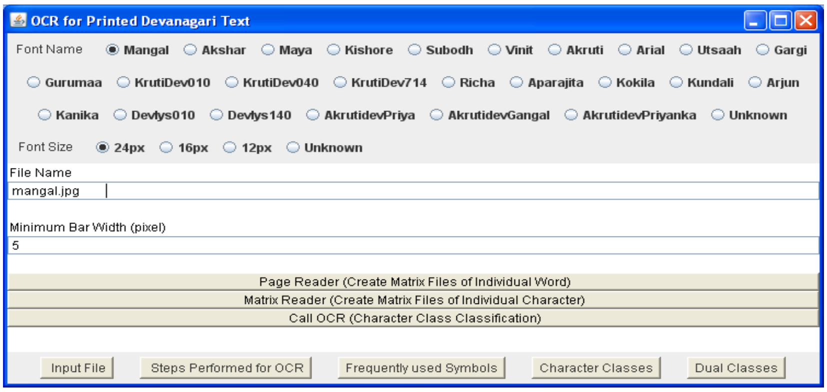


Figure 6. Screen Shot of the Main Menu

First, the program identifies the rows and then creates the binarized files of the each word enclosed in the rows. It performs four steps given in Table 10.

Table 10. Steps Performed to Create the Binarized Files of each Word

Step 1	Read the image file and convert into the binarized matrix
Step 2	Find out the start and end of rows by counting the highest number of 0 in row
Step 3	Find out the word in each start and end of each word in the row by counting the highest number of 0 in column
Step 4	Create the binarized file of each word

Then the program reads the binarized files of the each word and separates out the symbols in the middle zone. It performs eight steps as in Table 11.

Table 11. Steps Performed to find Symbols in Middle Zone from binarized File of each Word

Step 1	Read the binarized file of the word
Step 2	Remove white header horizontal line for correct image height of actual dark portion
Step 3	Remove white footer horizontal line for correct Image height of actual dark portion
Step 4	Remove white vertical line for correct image width of actual dark portion
Step 5	Get Shirorekha start row and shirorekha end row. (True if dark pixel width >= 60% of image height)
Step 6	Remove pixel above shirorekha i.e. the rows above the shirorekha start row.
Step7	Finding out the start and end of the symbols by counting the highest number of 0 in the column below shirorekha.
Step 8	Create the binarized file of each symbol

The system classifies the characters by computing various properties of each symbol (Table 12).

Table 12. Steps Performed to find Structural Properties in a given Character

Step 1	Read the binarized file of image
Step 2	Remove white header horizontal line for correct image height of actual dark portion
Step 3	Remove white footer horizontal line for correct Image height of actual dark portion
Step 4	Remove white vertical line for correct image width of actual dark portion
Step 5	Get Shirorekha (True if dark pixel width >= 60% of image height)
Step 6	Get All Bar (True if dark pixel width >= minimum bar width and dark pixel height >= 90% of image height)
Step 7	Merge bar if gap between bar is less than minimum bar width
Step 8	Ignore smaller bar if height is less than other bar
Step 9	Get All Bar Type (End/Mid/No Bar/Two Bar)
Step 10	Get All Shirorekha Touching

Some fonts are thin while others are thick. Accordingly minimum bar width parameter is set as given in Table 13.

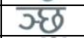
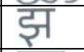
Table 13. Minimum Bar Width Parameter for different Fonts and Sizes

Font Name	**Minimum Bar Width**
Mangal (24px), GIST-DVOTAkshar (24px), GIST-DVOTSubodh, GIST-DVOTVineet, AkrutiDevYogini, Gargi, Kanika,	5
Arial Unicode MS, Kruti Dev 010, 040, 714, Mangal (16px, 12px) GIST-DVOTKishore,GIST-DVOTMaya(24px), Gurumaa, Utsaah, Richa, Aparajita, Kokila, devlys010, devlys140, AkrutiDevPriya, Arjun, AkrutiDevGangal, AkrutiDevPriyanka, Kundali,	3
GIST-DVOTAkshar(16,12px), GIST-DVOT Maya (16px)	2
GIST-DVOTMaya (12px)	1

It is also observed in the test data for these 25 fonts that there are various characters which fall in more than one character class due to difference in shape in different fonts. The exceptions found during the analysis over 25 fonts are shown in Table 14.

Table 14. Exceptions found during Analysis over 25 Fonts

Property	SNO	CHARACTER	MANGAL	AKSHAR	MAYA	KISHORE	SUBODH	VINIT	AKRUTI	ARIAL	UTSAAH	GARGI	GURUMAA	KRUTIDEV010	KRUTIDEV040	KRUTIDEV714	RICHA	APARAJITA	KOKILA	KUNDALI	ARJUN	KANIKA	DEVLYS010	DEVLYS140	AKRUTIDEVPRIYA	AKRUTIDEVGANGAL	AKRUTIDEVPRIYANKA
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
End Bar One Touching	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
End Bar Two Touching	1	ह्व	-	-	Y	Y	Y	Y	-	-	-	-	-	-	Y	Y	-	-	-	Y	Y	Y	Y	-	-	-	Y
	2	ह्ल	-	-	Y	Y	Y		Y	-	-	-	-	-	Y	Y	Y	-	-	Y	Y	Y	-	Y	-	-	Y
	3	ह्न	-	Y	Y	Y	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	4	ञ्ज	-	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	Y	Y	-	-	-	-	-	-	-	-
	5	ञ्च	-	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	6	भ	-	Y				Y			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	7	न्भ	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	8	ब्भ	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	9	भ्व	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	10	भ्र	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	11	ल्भ	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	12		-			Y				Y							Y		Y	Y							Y
End Bar Three Touching	1	ख्भ	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	2	भ्य	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	3	म्भ	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	4	अ	-		Y	Y	Y		Y		Y	Y	Y	Y	Y			Y		Y		Y			Y		Y
	5	श्र	-		-	-	-	-	-	-								Y									
	6	ल्श	-		-	-	-	-	-	-								Y									
	7		-		Y						-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
End Bar Four Touching	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
End Bar Five Touching	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mid Bar One Touching	1	क्क	-	Y		Y					Y	Y						Y	Y								
	2	क्व	-							Y	Y	Y															
	3	क्र	-							Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

	4		-		-	-	-	-	-	-	Y							Y	Y								
	5		-		-	-	-	-	-	-								Y	Y								
	6		-		-	-	-	-	-	-								Y	Y								
Mid Bar Two Touching	1		-		-	-	-	-	-	-	Y							Y	Y								
	2		Y		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	3		-				Y	Y			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mid Bar Three Touching	1		-		Y						-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
No Bar No Touching	1		-						Y		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
No Bar One Touching	1		-		-	-	-	-	-	-	Y							Y	Y								
	2		Y	Y	Y		Y				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	3		-	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y				Y	Y	Y	Y	Y	Y	Y
	4		-	Y	Y		Y								Y	Y		Y				Y		Y			
	5		-	Y	Y											Y											
No Bar Two Touching	1		Y		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	2		Y		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	3		-			Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y			Y		Y	Y	Y	Y	Y	
	4		-			Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y			Y		Y	Y	Y	Y	Y	
	5		-			Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y			Y		Y	Y	Y	Y	Y	
	6		Y		-	-	-	-	-	-											Y		Y				
	7		Y		-	-	-	-	-	-					Y												
No Bar Three Touching	1		-		-	-	-	-	-	-	Y																
	2		-							Y																	
	3		-	Y	-	-	-	-	-	-							Y								Y		Y
	4		-	Y	Y																						
Two Bar Two Touching	1		-	Y	Y		Y							Y	Y	Y							Y				
Two Bar Three Touching	1		-		Y			Y			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Two Bar Four Touching	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Two Bar Five Touching	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

An analysis of the occurrence of these exceptions is done on 5 samples downloaded from CIIL website [23]. The results of the analysis of the occurrence of these exceptions on 13113 words of 5 sample documents are shown in Table 15. The result shows that occurrence of these exceptions is very marginal.

Table 15. Analysis of occurrence of Exceptions

SNO	Symbol	No. of words in which symbol occurs	%age of occurrence	SNO	Symbol	No. of words in which symbol occurs	%age of occurrence
1	ह्व	-	0	21	म्प्ल	-	0
2	ह्ल	-	0	22	ङ्य	-	0
3	ह्न	-	0	23	ब्क	1	0.01
4	ञ्ज	-	0	24	क्ल	1	0.01
5	ञ्च	-	0	25	ह्म	1	0.01
6	न्भ	-	0	26	ल्श	3	0.02
7	ब्भ	-	0	27	क्क	3	0.02
8	भ्व	-	0	28	भ्र	4	0.03
9	ल्भ	-	0	29	म्भ	5	0.04
10	ख्भ	-	0	30	ल्ल	6	0.05
11	भ्य	-	0	31	च्छ	10	0.07
12	ष्ठ	-	0	32	झ	3	0.02
13	ब्क	-	0	33	झ	14	0.11
14	फ्ल	-	0	34	क्ष	25	0.19
15	ह्य	-	0	35	छ	42	0.32
16	ॐ	-	0	36	ष	88	0.67
17	प्ल	-	0	37	श	191	1.46
18	ल्ठ	-	0	38	भ	215	1.63
19	न्ठ	-	0	39	ज	274	2.09
20	ञ्छ	-	0	40	अ	330	2.52

7. CONCLUSIONS

On observing 469580 words from a variety of sources it is found that there are 345 symbols used more frequently in the middle zone. The frequency analysis done for two documents with different contents and sizes shows that 97% text is covered by the single characters and the presence of the conjunct characters in these documents is around 3% only which is very less in comparison to the occurrences of single characters. This is despite the fact that the numbers of possible conjunct symbols are much higher than the possible number of single characters. The overall coverage by these identified 345 symbols is found to be 99.97% on 13113 words in 5 sample-documents from CIIL Text Corpora.

There are 16 different character classes to which a character may belong to based on the structural properties that are invariant across fonts and sizes. A detailed classification scheme is developed to identify the symbols in each of these classes. Some exceptions are found during the analysis over 25 fonts where a symbol falls in another class due to differences in shapes of different fonts. The occurrences of these exceptions from 13113 words in 5 sample-documents from CIIL Text Corpora are found to be very marginal.

The two main sources of errors are incorrect segmentation and classification. To enhance the accuracy rate of character recognition, first level of classification can be done using these 16 structural features which are invariant across fonts and sizes. The second level of classification using any other feature extraction technique can be done on a class which is not further segregated.

The utility of the proposed classification approach is to enhance the recognition accuracies by reducing the errors introduced due to font and size variations.

ACKNOWLEDGEMENTS

I thanks to Dr. C. Vasantha Lakshmi and Dr. M. Hanmandlu for their supervision and guidance. I thanks to Dr. C. Patvardhan for his timely and valuable suggestions.

REFERENCES

[1] A. N. Holambe, S. N. Holambe, R. C. Thool, "Comparative study of Devanagari handwritten and printed character & numerals recognition using Nearest-Neighbor classifiers" in 3rd IEEE International Conference on Computer Science and IT (ICCSIT), 2010 pp. 426-430
[2] Suryaprakash Kompalli, Sankalp Nayak, Srirangaraj Setlur, Venu Govindaraju, "Challenges in OCR of Devanagari documents" in Eighth International Conference on Document Analysis and Recognition, 2005 Proceedings. pp. 1-5
[3] S. Kompalli, S. Setlur, V. Govindaraju, "Design and comparison of segmentation driven and recognition driven Devanagari OCR" in Second International Conference on Document Image Analysis for Libraries, 2006. DIAL '06, pp. 1-7
[4] B. B. Chaudhuri, U. Pal, "An OCR system to read two Indian language scripts: Bangla and Devanagari (Hindi)" in Fourth International Conference on Document Analysis and Recognition, 1997 Proceedings pp. 1011-1015
[5] Veena Bansal, R. M. K. Sinha, "Integrating Knowledge Sources in Devanagari Text", IEEE Trans. on Systems, Man, and Cybernetics-Part A: Systems and Humans, Vol. 30, No. 4, pp. 500-505, July 2000.
[6] Pooja Agrawal, M. Hanmandlu, Brejesh Lall, "Coarse Classification of Handwritten Hindi Characters", International Journal of Advanced Science and Technology, Vol. 10, pp. 43-53, Sep. 2009.
[7] C. Vasantha Lakshmi, C. Patvardhan, "A high accuracy OCR system for printed Telugu text" in Conference on Convergent Technologies for Asia-Pacific Region TENCON 2003, p725-729
[8] Hindi Magazine, Sukravar, 18 Dec - 24 Dec, 2010
[9] Hindi Magazine, India Today, 17 Mar. - 23 Mar., 2011
[10] Hindi News Paper, Sunday Navbharat Times, 10 Nov. 2013
[11] Hindi News Paper, Rashtriya Sahara, 14 Nov. 2013
[12] Vishnu Sharma, Panchtantra Ki 101 Kahaniya,, 7th ed., Manoj Publications, Delhi, 2010
[13] Vivek Mohan, Tenaliram Ki Kathain, Jhole Main Katora, 1st ed., Raja Pocket Books,Delhi, 2004
[14] Hindi Nibandh Aevam Patra Lekhan, Top publication, Delhi
[15] Dr. Om Prakash Ji Maharaj, Ayurved Ka Chamatkar Haldi, Laxmi Prakashan, Delhi
[16] Hawkins Akhil Bhartiya Pak Pustika, Hawkins Cookers Limited, Bombay
[17] Manju Mehrotra, Manak Vyakaran Vyavahar, Part-I, Prabhat Publicating House, New Delhi
[18] Jagdish Singh, Gyan Sagar 6, Gita Publishing House, Delhi
[19] Jagdish Singh, Gyan Sagar 3, Gita Publishing House, Delhi
[20] Dr. Sukhpal Singh, Satrangi, 1st ed., Saviyo Publications Pvt. Ltd., New Delhi, 2009
[21] Hindi Magazine, Sukravar, 18 Dec. - 24 Dec., 2010, pp. 54-55
[22] Hindi Magazine, Sukravar, 18 Dec. - 24 Dec., 2010, pp. 58
[23] (2014) Indian Languages Corpora on CIIL website. [online]. Available:
http://www.ciilcorpora.net/hindisam.htm,
http://www.ciilcorpora.net/pdf/hindipdf/HINDI1.pdf,
http://www.ciilcorpora.net/pdf/hindipdf/HINDI2.pdf,
http://www.ciilcorpora.net/pdf/hindipdf/HINDI3.pdf,
http://www.ciilcorpora.net/pdf/hindipdf/HINDI4.pdf,
http://www.ciilcorpora.net/pdf/hindipdf/HINDI5.pdf

AUTHORS

Manoj Kumar Gupta received his M.Sc. degree in Statistics in 1989 from Agra University. He received P. G. Diploma in Computer Science and Application in 1990 and M. Phil. degree in Computer Science in 2010 from Dayalbagh Educational Institute, Dayalbagh, Agra. Currently He is working in National Informatics Centre, New Delhi as Scientist-E and also pursuing the Ph.D. (C.S.) from Dayalbagh Education Institute, Dayalbagh, Agra. He is having an experience of 23 years in Application Software Development.

C. Vasantha Lakshmi is working as an Associate Professor in the Department of Physics and Computer Science, Dayalbagh Educational Institute, Agra. She has an M. Tech. in Computer Science from University of Hyderabad, Hyderabad and Ph.D. in the area of OCR for printed Telugu text from the Dayalbagh Educational Institute. She was awarded the ISCA Young Scientist Award for this work. She has also proved her research credentials by publishing a number of papers in various Journals and Conference proceedings and received several awards. Her research interests are broadly in the area of Pattern recognition, Image Processing, Soft computing and evolutionary optimization She has handled several research projects both minor and major from several funding agencies like UGC, Department of Information Technology etc.

Madasu Hanmandlu received the B. E. degree in Electrical Engineering from Osmania University, Hyderabad, India, in1973, the M. Tech. degree in power systems from R.E.C. Warangal, Jawaharlal Nehru Technological University (JNTU), India, in1976, and the Ph. D. degree in control systems from Indian Institute of Technology, Delhi, India, in 1981. From 1980 to1982, he was a Senior Scientific Officer in Applied Systems Research Program (ASRP) of the Department of Electrical Engineering, IIT Delhi. He joined the EE department as a lecturer in1982 and became Assistant Professor in1990, an Associate Professor in1995 and finally a Professor in1997. He was with Machine Vision Group, City University, London, from April–November, 1988, and Robotics Research Group, Oxford University, Oxford from March–June, 1993, as part of the Indo-UK research collaboration. He was a Visiting Professor with the Faculty of Engineering (FOE), Multimedia University, Malaysia from March 2001 to March 2003. He worked in the areas of Power Systems, Control, Robotics and Computer Vision, before shifting to fuzzy theory. His current research interests mainly include Fuzzy Modelling for Dynamic Systems and applications of Fuzzy logic to Image Processing, Document Processing, Medical Imaging, Multimodal Biometrics, Surveillance and Intelligent Control. He has authored a book on Computer Graphics in 2005 under PBP publications and also has well over 220 publications in both conferences and journals to his credit. He has guided 20 Ph.D. and120 M. Tech. students. He has handled several sponsored projects. He was an Associate Editor of both Pattern Recognition Journal (January 2005–March 2011) and of IEEE Transactions on Fuzzy Systems and are viewer to other journals such as Pattern Recognition Letters, IEEE Transactions on Image Processing and Systems, Man and Cybernetics(January 2007–2010). He is a senior member of IEEE and is listed in Reference Asia; Asia's who's who of Men and Women of achievement; 5000 Personalities of the World (1998), American Biographical Institute. He was a Guest Editor of Defence Science Journal for the special issue on "Information sets and Information Processing" September, 2011.

C. Patvardhan is working in the Dayalbagh Educational Institute, Agra as Professor in Electrical Engineering and Coordinator of the PG programme in Faculty of Engineering, DEI. He obtained his BSc (Engg.) degree from Dayalbagh Educational Institute in 1987, M. Tech. from IISc, Bangalore in Computer Science in 1989 and his PhD in 1994 from Dayalbagh Educational Institute.

He has published more than 250 papers in Journals and Proceedings of Conferences and has won 15 Best Paper / Best Presentation Awards. He has also published one book and has been an Editor of three Conference proceedings. He has been the Investigator / Co-Investigator of several funded R&D projects including a DST-DFG Indo-German collaborative project. His current research interests are Quantum and Soft Computing and Image Processing. He has been a training consultant on Advanced Algorithms to Cadence Design Systems and Atrenta Communications. He has delivered more than 70 invited talks in national and international meetings.

He is a life member of Computer Society of India, Systems Society of India and Indian Science Congress Association and a Fellow of United Writers Association of India, Chennai..