International Journal Of

Recent Scientific Research

ISSN: 0976-3031

Volume: 7(2) February -2016

CLASSIFICATION OF DEVANAGARI CHARACTER BASED ON LEFT SURFACE CAVITY

Manoj Kumar Gupta., Vasantha Lakshmi C and Patvardhan C

THE OFFICIAL PUBLICATION OF
INTERNATIONAL JOURNAL OF RECENT SCIENTIFIC RESEARCH (IJRSR)
http://www.recentscientific.com/ recentscientific@gmail.com

ISSN: 0976-3031

Available Online at http://www.recentscientific.com

International Journal of Recent Scientific Research
Vol. 7, Issue, 2, pp. 8741-8746, February, 2016

International Journal of Recent Scientific Research

RESEARCH ARTICLE

CLASSIFICATION OF DEVANAGARI CHARACTER BASED ON LEFT SURFACE CAVITY

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

[1]Department of Physics and Computer Science, Dayalbagh Educational Institute, Dayalbagh, Agra-282005, India
[2]Department of Electrical Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra-282005, India

ARTICLE INFO

Article History:

Received 15th September, 2015
Received in revised form 21st November, 2015
Accepted 06th January, 2016
Published online 28th February, 2016

Key words:

Optical Character Recognition, Conjunct Character, Water Bodies, Left Surface Cavity

ABSTRACT

Size of the character set with which it has to deal is the main area of concern during development of any OCR. The presence of larger number of character (theoretically 46656) in the middle zone of the Devanagari Text makes the development of OCR for Devanagari complex. Hence every effort is required to break this larger character set into small manageable chunks so that recognizer has only to deal with a small subset of symbols. The identification of various types of structural properties which are invariant to fonts and sizes reduces the bigger character set of Devanagari Text into small manageable chunks. This set can be reduced further in each class if more structural properties are identified. This results in development of a recognizer with increased accuracy and better response time. One such structural property is the number of cavities present in the left surface of the character. This structural property further reduces the character set identified by the various other structural properties viz. bar type, number of shirorekha touching and water bodies.

INTRODUCTION

Akshara is the single phonetic unit in the Devanagari script which has three strips or zone viz. upper zone, middle zone and lower zone. The middle zone constitutes the core character while upper and lower zone is for modifier i.e. matra. The initial separation of symbols is done based on the symbols present in various zones and the complexity is reduced by separation of core characters in the middle zone and modifiers.

There are various single and conjunct characters present in the middle zone. The connected component is treated as one symbol for the identification of possible list of characters in the middle zone. Though theoretically there can be 46656 conjunct characters, study on 469580 words from a variety of sources shows that there are only 345 frequently used symbols in the middle zone (Gupta *et al*, 2014). This reduces the middle zone character set to 345 symbols.

Classification of characters can be done by feature extraction. The various feature extraction methods used by the past researcher includes shape based features, statistical features (Bansal and Sinha, 2000), gradient and curvature based feature extraction methods (Kompalli *et al*, 2005) and features extracted from the concept of water reservoir (Pal and Roy, 2004)..

For consistent and higher accuracy, a classification scheme needs to take care into account that further reduction in character set needs to be done based on the identification of various structural properties which are invariant to font and sizes as the Devanagari character shape changes drastically with fonts.

There are various structural properties or distinguishable features which remain same with various fonts and sizes. These include presence or absence of vertical bar, number of places touching to shirorekha. This scheme reduces 345 frequently used characters into 16 small manageable classes (Gupta *et al*, 2014).

Further reduction of character set is done on the basis of water bodies (Gupta, Vasantha and Patvardhan, 2016) which are found after darkening the white pixels which are NOT visible from bottom as well as darkening the white pixels which are visible from left and right. This classification is done based on

Corresponding author:* **Manoj Kumar Gupta
Department of Physics and Computer Science, Dayalbagh Educational Institute, Dayalbagh, Agra-282005, India

number of water bodies found in different symbols and is a good basis for classification which is invariant to font and size. Whereas the vertical bar and touching count property reduces the character set, the number of water bodies separate out the single and conjunct character. This helps in enhancing the overall recognition accuracy as the single character covers 97% text and the possible number of single characters is much less in comparison of number of conjunct character (Gupta *et al*, 2014).

Since each word contains large numbers of components or symbols hence while component level recognizers perform well but in practical situation, the word level and document level accuracies are not acceptable (Rasagna *et al*, 2009). Enhancement in recognition accuracies can also be achieved by identifying the possible character set for single letter words.

The study of various words from various sources shows that there are only 17 characters in the middle zone which are occurring in single letter words with 50% coverage of single letter words by four characters क ह थ म (Gupta *et al*, 2016).

***Motivation for the present work*:** Size of the character set with whom it has to deal is the main area of concern and hence every effort is required to break this larger character set into small manageable chunks so that recognizer looks only the small number of symbols. There will be small number of characters in each class if more and more structural properties are identified. Whether any more structural property can be identified which can further reduce the size of character set so that recognizer with higher accuracy and better response time can be developed? This is the pertinent question behind the motivation for undertaking this research work.

MATERIALS AND METHODS

As for identifying the shape of any object by closed eyes, one will identify the shape of an object by touching it. In doing so he tries to identify the hole or cavity present in the object along with convexity/concavity of it. After sensing this information, he identifies the shape of the object.

Similarly the Devanagari character can be classified on the basis of number of cavity present in the left surface of the character. A Devanagari character contains various curves, holes and small connecting horizontal and vertical lines apart from Bar and Shirorekha. Also the text in Devanagari is read from left to right hence the number of left surface cavity can be a good basis for classifying the characters. The Image of characters with an arrow showing the left surface cavity is shown below in Figure 1.

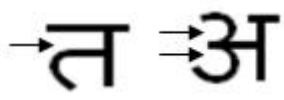

Figure 1 Image of characters with an arrow showing the Left Surface Cavity

Various steps required to be performed starting from reading a page of text to classify a character are given below in Figure 2.

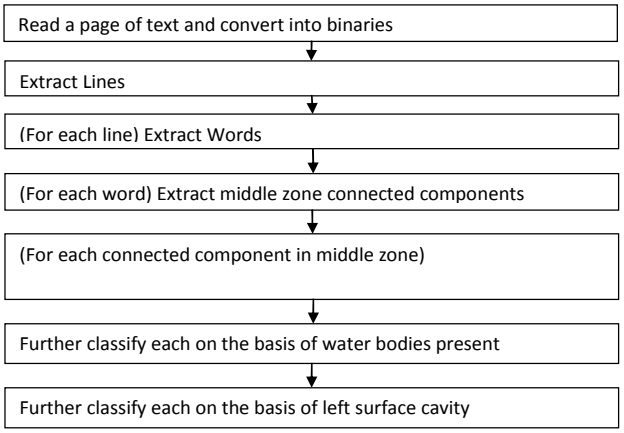

Figure 2 Steps performed to classify a character on the basis of structural properties

To correctly identify the cavities, it is necessary to perform the equalization of the shapes from all side except the cavity i.e. till the beginning of the cavity. Various steps performed for each connected component in middle zone to identify the number of left surface cavity are given below in Figure 3.

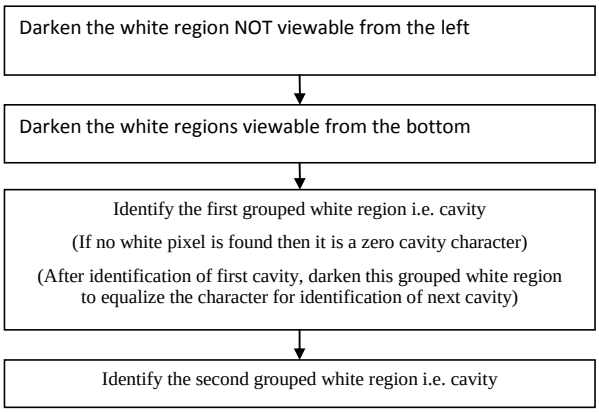

Figure 3 Steps performed to identify the number of left surface cavity present in a character

Step I (Darken the white regions NOT viewable from the Left i.e. right side equalization of the character)

Travers the binarized matrix for each row of the symbol from top to bottom and column from left to right till dark pixel is found. When dark pixel is found then make all the pixels right to this pixel as dark. Image of binarized file of Symbol अ after darkening the white region NOT viewable from left is shown in Figure 5 and steps are shown below in Figure 4.

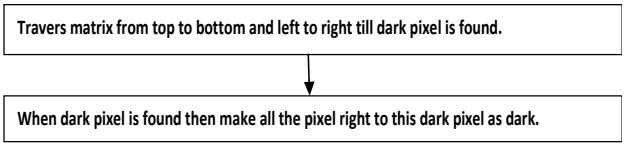

Figure 4 Steps to darken the white regions NOT viewable from the left

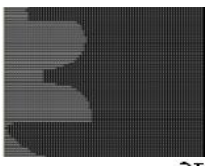

Figure 5 Image of binarized file of Symbol अ after darkening the white region NOT viewable from the left

Table 1 Classification of total 345 characters based on bar type, touching count, water bodies and left surface cavity

SNO	Property (Bar Type) (Touching Count)	No. of char	Property (Water Bodies)	No. of char	Property (Left Surface Cavity)	No. of Char	Character
1	End Bar One Touching	49	Less than two water bodies	15	Zero cavity	1	T
					One cavity	6	ग त न ब व ↗
					Two cavity	8	च ज ञ त्र ज्ञ त्त ब्र व्र
			Two or more water bodies	34	One cavity	24	त्न त्ज न्ज न्त्त न्न ब्ज ब्त ब्ल ल्ज ल्ल व्ल त्र त्व न्व न्त्र ब्व ल्व ल्व व्व न्त्व ल्च ल्त त्ल न्च
					Two cavity	10	ज्ज न्व ज्व ज्ज्व ञ्ज ञ्च च्च च्व त्व ज्र
2	End Bar Two Touching	70	Less than two water bodies	13	Zero cavity	3	प ष प्त
					One cavity	7	म स य ख घ ग्र प्र
					Two cavity	3	अ झ घ्र
			Two or more water bodies	57	Zero cavity	4	ष्व ष्च प्ल प्न
					One cavity	42	स म्र ख्त म्त म्न म्ल स्त स्न स्ल ख्व म्व स्च स्ज स्व स्व स्न स्त्व त्म त्स न्म न्स ब्म ब्स ल्म ल्स त्प त्य न्य ब्य ल्प ल्य व्य न्त्य न्त्य त्व्य म्य म्ज त्स्न त्प्र ब्ल्य घ्न घ्व
					Two cavity	11	ज्झ ज्म ञ्झ च्य ज्य ञ्य ञ्ज्य ह्य ह्म ग्र म्र
3	End Bar Three Touching	54	Less than two water bodies	6	One cavity	4	ध भ थ भ्र
					Two cavity	2	श्र ध्र
			Two or more water bodies	48	Zero cavity	6	प्स ष्म प्प प्य ष्प ष्य
					One cavity	40	ल्म्स त्क्ष न्भ ब्भ ख्म ख्स म्म म्स स्म स्स त्स्य त्स्प त्थ न्थ न्ध ब्थ ब्ध ल्थ ल्भ ख्य म्य म्प य्य स्ख स्प स्य स्त्य स्व न्ध स्प्र म्प्ल ष्प्र श्ध श्च ध्न घ्य ध्व थ्व भ्व क्ष
					Two cavity	2	श्र म्प्र
4	End Bar Four Touching	13			One cavity	13	ख्भ म्भ म्थ स्थ न्ध्य म्प्य क्ष्म क्ष्य भ्य थ्य ध्य थ्स ध्म
5	End Bar Five Touching	1			One cavity	1	स्थ्य
6	Mid Bar One Touching	11			One cavity	7	क त्क न्क ल्क ल्क क ङ
					Two cavity	4	ऋ त्त च्क क्र
7	Mid Bar Two Touching	17			Zero cavity	3	फ ष्क फ
					One cavity	13	क्ट त्फ ल्फ म्क स्क न्प क्स क्ड स्क्र म्प फ्र ष्क्र न्फ्र
					Two cavity	1	ङ्ग
8	Mid Bar Three Touching	8			Zero cavity	2	फ्ट ष्फ
					One cavity	6	क्श म्फ स्फ क्स्ड क्स्ट म्फ्र
9	No Bar No Touching	17			One cavity	12	ट ट ढ ल ट ज ञ न्ट न्ड ट्न ब्ट :
					Two cavity	5	ट ञ ञ ञ्ज ट
10	No Bar One Touching	53	Less than two water bodies	35	Zero cavity	2	ट ड
					One cavity	22	ट ठ ढ र ख ट स्ट द ट्ट द्र ळ रु न्त त्न्ड त्स्ट स्ट स्न न्स्ट ब्ट ट छ
					Two cavity	11	इ उ ड ऽ ऊ ह ढ्ढ द्व झ द ॐ
			Two or more water bodies	18	One cavity	11	न्द न्ह ब्द ल्द ल्ह ल्भ व्ह ल्ग रू ल्ट ल्ड
					Two cavity	7	च्छ ञ्छ ह्व ह्ल ह्न ह्न न्द्र
11	No Bar Two Touching	22	Less than two water bodies	17	Zero cavity	6	ए प ट ष्ड ष्ठ ष्ट
					One cavity	11	त्श म्ड स्ट न्श स्रु न्ह स्ड श श ६ क्ष
			Two or more water bodies	5	One cavity	5	म्द म्ह न्स्ट न्स्ड ल्श
12	No Bar Three Touching	4			Zero cavity	2	प्श ष्ए
					One cavity	2	ख्श स्श
13	Two Bar Two Touching	7			One cavity	7	क्च क्न क्ल क्व क्क क्त्र क्ज
14	Two Bar Three Touching	11			Zero cavity	4	फ्त फ्ल फ्क फ्ज
					One cavity	7	क्ख क्प क्म क्य क्स क्फ ल्क्य
15	Two Bar Four Touching	6			Zero cavity	4	फ्म फ्य फ्स फ्फ
					One cavity	2	क्थ न्फ्य
16	Two Bar Five Touching	2			Zero cavity	1	फ्थ
					One cavity	1	क्स्थ

Step II (Darken the white regions viewable from the bottom i.e. left side equalization of the character)

Now Travers the binarized matrix for each row of the symbol in the reverse order i.e. from bottom to top and column from left to right till dark pixel is found. When dark pixel is found then make all the pixel below to this dark pixel as dark. Image of binarized file of Symbol अ after darkening the white region viewable from bottom is shown in Figure 7 and steps are shown in Figure 6.

Step II (Darken the white regions viewable from the bottom i.e. left side equalization of the character)

Now Travers the binarized matrix for each row of the symbol in the reverse order i.e. from bottom to top and column from left to right till dark pixel is found. When dark pixel is found then make all the pixel below to this dark pixel as dark. Image of binarized file of Symbol अ after darkening the white region viewable from bottom is shown in Figure 7 and steps are shown in Figure 6.

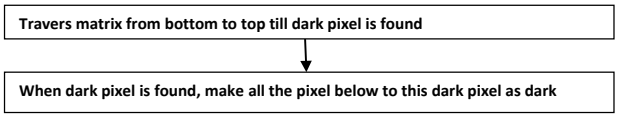

Figure 6 Steps to darken the white regions viewable from the bottom

Figure 7 Image of binarized file of Symbol अ after darkening the white region viewable from bottom

Step III (Identify the first grouped white region i.e. cavity)

Now Travers the binarized matrix for each row of the symbol in the reverse order i.e. from bottom to top and column from left to right till white pixel is found. When white pixel is found then make all the white pixel above to this white pixel as dark. Image of binarized file of Symbol अ after identifying the first grouped white region i.e. cavity is shown in Figure 9 and steps are shown in Figure 8. If no grouped white pixel is found then it is a zero cavity character. After identification of first cavity, darken this grouped white region to equalize the character for identification of next cavity.

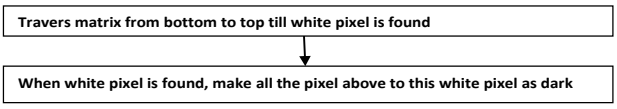

Figure 8 Steps to identify the first cavity

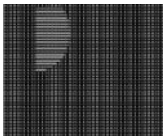

Figure 9 Image of binarized file of Symbol अ after identifying the first cavity

Step IV (Identify the second grouped white region i.e. cavity)

Repeat the steps mentioned in steps III once again for identifying the second grouped white region i.e. cavity. Image of binarized file of Symbol अ after identifying the second grouped white region i.e. cavity is shown in Figure 10

Figure 10 Image of binarized file of Symbol अ after identifying second cavity

Image of the characters after completion of every step during identification of cavity is shown below in Figure 11.

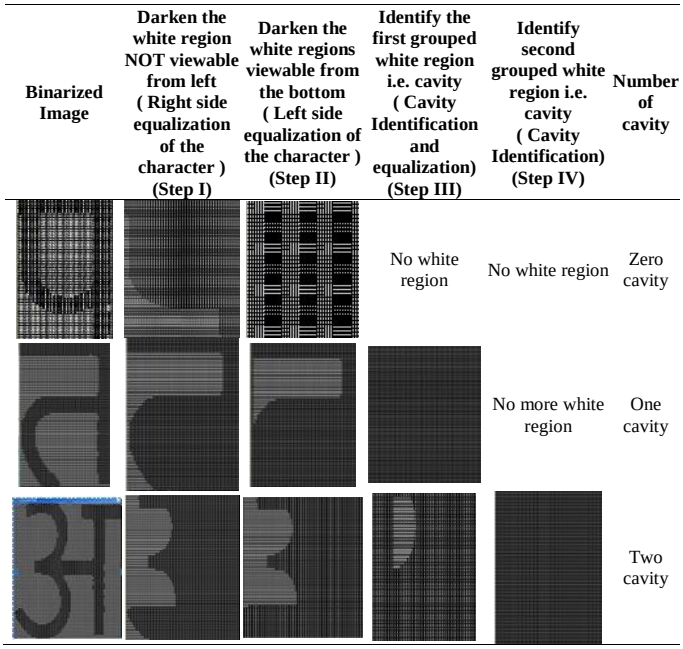

Figure 11 Image of the characters after completion of every step during identification of cavity

RESULTS

Classification of 345 frequently used Devanagari characters

The number of cavity present in the left surface of the character further reduces the classes identified by the various other structural properties viz. bar type, number of shirorekha touching and water bodies. The detailed classification of total 345 characters based on bar type, touching count, water bodies and left surface cavity is shown below in Table 1.

Pattern of Left Surface Cavity in Single Character

The pattern of left surface cavity in single character is summarize and given below in Table 2.

Table 2 Pattern of left surface cavity in single character

SNO	Property (Left Surface Cavity)	No. of Char	Single Character
1	Zero Cavity	6	ए प फ ष ण ट
2	One Cavity	23	क ख ग घ छ ट ठ ढ त थ द ध न ब भ म य र ल व श स क्ष
3	Two Cavity	14	अ इ उ ऊ ऋ ड च ज झ ह ञ त्र ज्ञ श्र

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 3.

Table 3 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			

Figure 12 Image of Test Data File of Mangal Font with 24px Font Size

Table 4 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
One Cavity	1	ज	-	-	-	-	-	-	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	-	Y	Y	-	-	-	-	-	-
	2	च	-	-	-	Y	-	-	-	Y	Y	-	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	Y
	3	ए	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	4	ष	Y	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-	-	-	Y	-	-	-	-	-
	5	ष	Y	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-	-	-	Y	-	-	-	-	-
	6	स	Y	-	-	-	-	-	-	Y	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Two Cavity	1	क्ष	-	Y	-	-	Y	-	-	Y	Y	Y	-	-	-	-	-	Y	Y	-	-	-	-	-	-	-	-
	2	त्र	-	Y	Y	Y	Y	Y	-	Y	Y	Y	Y	-	-	-	-	Y	Y	-	-	-	-	-	-	-	-
	3	ज्ञ	-	Y	Y	Y	Y	Y	-	-	Y	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-
	4	श्र	-	Y	Y	-	Y	Y	Y	Y	-	-	Y	Y	-	Y	Y	Y	Y	-	Y	-	Y	-	Y	-	Y
	5	छ्य	-	Y	-	-	Y	-	-	Y	-	-	Y	-	-	Y	-	Y	-	-	-	-	Y	Y	-	Y	Y
	6	ख्य	-	-	Y	-	Y	-	-	-	Y	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-
	7	थ्य	-	-	-	Y	Y	Y	-	-	-	-	-	-	-	-	-	Y-	-	-	-	-	-	-	-	-	-
	8	भ्य	-	-	-	Y	Y	Y	-	-	-	-	-	-	-	-	-	Y	-	-	-	-	-	Y	-	-	-
	9	घ्य	-	-	-	Y	Y	Y	-	-	-	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-
	10	क्ष	-	-	-	Y	Y	Y	-	-	-	-	-	-	-	-	-	Y	-	-	-	-	-	Y	-	-	-

The Figure 12 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.

Exception found during analysis over 25 fonts

It is also observed in the test data for these 25 fonts that there are various characters which fall in other character class. The exceptions are shown in Table 4.

DISCUSSION

The complexity of presence of larger number of character in the middle zone of the Devanagari Text is reduces to the small manageable chunks by classifying the character using the structural properties. The result shows very less number of characters in most of the classes. The character shows the presence of zero cavity in 6 single characters, one cavity in 23 single characters and two cavity in 14 single characters. The conjunct character starting with these single characters also exhibit the same number of cavity.

CONCLUSION

During the OCR process, the unknown character symbol can be processed to first identify the class it belongs to and then a simple recognizer can be used to recognize it. Since each recognizer would have to recognize the unknown symbol out of a much smaller number of possibilities, it would be much easier to design these. Thus, this classification step can be used with advantage along with any OCR system as a preprocessor block.

References

Gupta M. K., Vasantha C., Hanmandlu M., Patvardhan C. (2014). An Exhaustive Font and Size Invariant Classification Scheme for OCR of Devanagari Character. *International Journal on Natural Language Computing*, 4(1): 1 - 21

Gupta M. K., Vasantha C., Patvardhan C. (2016). Classification of Devanagari Characters based on Water Bodies. *International Journal of Computer & Mathematical Sciences*, 5(1): 18 - 27

Gupta M. K., Vasantha C., Patvardhan C. (2016). Identification of Character Pattern in Devanagari Words for Enhancement of Recognition Accuracy. Advances in Computer Science and Information Technology, 3(1): 5 - 8

Rasagna V., Kumar A., Jawahar C. V., Manmatha R. (2009). Robust Recognition of Documents by Fusing Results of Word Clusters. International Conference on Document Analysis and Recognition. 566 - 570

Bansal Veena, Sinha R. M. K. (2000). Integrating Knowledge Sources in Devanagari Text. IEEE Trans. on Systems, Man, and Cybernetics-Part A: Systems and Humans, 30(4) 500 - 505

Kompalli Suryaprakash, Nayak Sankalp, Setlur Srirangaraj, Govindaraju Venu (2005). Challenges in OCR of Devanagari documents. Eighth International Conference on Document Analysis and Recognition Proceedings. 1 - 5

Pal U., Roy P. P. (2004). Multi oriented and Curved Text Lines Extraction from Indian Documents. IEEE Transaction on Systems, Man and Cybernetics-Part B: Cybernetics, 34(4): 1676 - 1684

****

How to cite this article:

Manoj Kumar Gupta *et al.*2016, Classification of Devanagari Character Based on Left Surface Cavity. *Int J Recent Sci Res.* 7(2), pp. 8741-8746.

TSSN 0976-3031