# Optimizing Character Class Count for Devnagari Optical Character Recognition

Jasbir Singh and Gurpreet Singh Lehal
Department of Computer science, Punjabi University, Patiala, India
jbs.5@rediffmail.com, gslehal@gmail.com

**Abstract.** Optical character recognition is a widely used technique for generating digital counterpart of printed or handwritten text. A lot of work has been done in the field of character recognition of Devnagari script. Devnagari script consists of several basic characters, half form of characters, vowel-modifiers and diacritics. From character recognition point of view only 78 character classes are sufficient for the identification of these characters. But in Devnagari the characters fuse with each other, which result in segmentation errors. Therefore to avoid such errors we shall consider such compound characters as separate recognizable unit. We have identified 864 such compound characters that make a total of 942 recognizable units. But it is very difficult to handle such a large number of classes; therefore we have optimized the character class count. We have found that the first 100 classes can contribute to 98.0898% of the overall recognition.

**Keywords**: Conjuncts, Segmentation, Recognizable unit.

## Introduction

Optical character recognition belongs to the family of techniques performing automatic identification. It deals with the problem of recognizing optically processed characters. The character recognition work on Devnagari script started in 70's when Sinha and Mahabala [1] presented a syntactic pattern analysis system with an embedded picture language. Sinha and Bansal [2] have discussed the use of various knowledge sources at all levels in Devnagari document processing system. Chaudhuri and Pal [3] have suggested primary grouping of characters, where each character is assigned to one of the three groups namely basic, modifier and compound character group before going for actual recognition process. Bansal and Sinha [4-5] have presented method for segmentation and decomposition of Devnagari composite characters into their constituent symbols. Kompalli et al [6] have discussed the wide range of challenges in Devnagari script that are not seen in Latin based scripts. They also mentioned that half consonants have different shapes from full-consonants; therefore the use of post-processing techniques or half-consonant classifiers for the left part can improve conjunct recognition. In our work we shall consider each compound character as separate class so as to reduce the errors introduced due to over and under segmentation of such characters. Therefore in this paper we have presented the analysis and optimization of characters classes that would be sufficient to get the desired recognition rate.

## Problems with segmentation (Need of multiple classes)

One of the significant phases in any optical character recognition system, upon which the performance of the overall system depends is the *segmentation phase*. Segmentation phase consists of the line segmentation, word segmentation and character segmentation. Out of these, character segmentation is most critical one, as the most of the recognition errors in optical recognition system are due to the character segmentation. In Devnagari the constituent characters may join horizontally or vertically to form new characters. When the constituent characters (*consonants*) fuse horizontally (*laterally*), they result in the formation of conjuncts (Fig. 1).

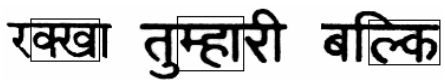



Fig. 1.

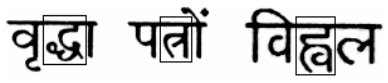



Fig. 2.

Although in the resulting conjuncts, the constituent characters are at adjacent positions, but they fuse laterally in such a way that there is no vertical space between them and hence it becomes very difficult to separate them during segmentation phase. In certain cases the constituent characters may combine in such a way which leads to the formation of new single character in which the constituent characters does not appear at adjacent positions i.e. they merge with each other (Fig. 2)

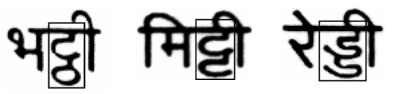
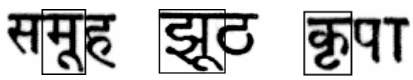



Fig. 3.




(a) (b) (c)

Fig. 4.

Similarly the constituent components may combine vertically, which result in the formation of characters having more height than the normal character height (Fig. 3).There is another class of characters in which height of the character itself or when combined with descenders (lower zone vowel modifiers) is such that it results in the problems in segmentation. In some cases the height of the primitive character is so large as compared to the adjacent consonant-descender combination that it leads to under segmentation (Fig. 4-a). Similarly the height of consonant itself may be so large that it will lead to the over segmentation of consonant-descender combination (Fig. 4-b). In few cases the consonant-descender combination itself is small in height as compared to adjacent character that it lead to the under segmentation (Fig. 4-c). From above it is clear that it is quite difficult to separate the constituent characters from these compound characters. If we try to segment them, it will lead to over or under segmentation and hence segmentation errors. Therefore treating such type of compound characters as a separate recognizable unit can decrease the segmentation problems. In the same way all the consonant-descender combinations will be treated as separate classes.

## Identification/Optimization of the classes

Before getting into the optical recognition process one must know the character classes that are going to be used for the underlying script. Any word in Devnagari can be divided into three zones: middle, upper and lower zone. The table shown below (Table 1.) provides some of the recognizable units corresponding to all three zones. The recognizable unit is the smallest possible unit that can be recognized by character recognition process. For example ग contribute to both ग and ा, which are two recognizable units that will be separated during segmentation phase if one removes the headline. Some other examples are given after Table 1. Apart from these basic recognizable units there exist a large number of compound characters that can be treated as separate recognizable units. Therefore we have to identify them, as they will contribute to the recognition process. As stated above there are four different categories of the character combination that are to be identified:

*Category-1:* The combination in which consonants fuse to create conjuncts (Fig-1).

*Category-2:* The combination in which the adjacent consonants fuse to form a composite character having more height as compared to other (Fig-3).

*Category-3*: The characters may join which result in a single character (Fig-2).

*Category-4:* All of the consonant with the lower vowel modifier (Fig-4).

In order to identify the possible character classes a corpus of approximately 3 million words has been used. The corpus comprises of Unicode data, therefore most of character combinations corresponding to categories 1-3 are identified with the help of diacritic ् (*halant)*. It is to be noted that in many cases even the presence of halant does not cause the adjacent consonants to combine to form the compound character corresponding to categories 1-3. For example in the word ट्वेंटी, ट and व does not result in conjunct despite the presence of halant. Therefore such characters along with the exceptions, which do not form conjuncts even in the presence of halant are identified. For example ट will not form the compound characters corresponding to categories 1-3 with any consonant except with ट ठ य र. For the category 4, the presence of descender is checked after the consonant. Again there are certain exceptions to it, for example in the word अद्भुत even though द and भ can lead to conjunct द्भ, but the presence of ु will result in separate consonant with descenders द् and भु. Therefore care is taken to count them separately.

**Table 1.** First 10 recognizable units (out of 78) along with their frequency of occurrence.

| S.No | Recognizable unit | Overall % occurrence |
|---|---|---|
| 1. | ा | 20.7631 |
| 2. | े | 7.4461 |
| 3. | क | 5.2938 |
| 4. | र | 5.2631 |
| 5. | ी | 3.7936 |
| 6. | न | 3.7870 |
| 7. | स | 3.4365 |
| 8. | ह | 3.3931 |
| 9. | म | 3.2991 |
| 10. | ि | 3.2023 |

In the above table आ, ऑ, ओ, ओ, औ, ि, ी, ॉ, ॉ, ो, ो, ौ, ग, ण, श, ग़ contribute to ा. Similarly "॥ Contribute twice to ।", "ऐ, ओ, ो contribute to े", "औ, ौ contribute to ै", "ऍ, ऑ, ॉ contribute to ॅ" and "े, ऐ, ओ, ॉ contribute to े".
The following table (Table 2.) depicts first few recognizable units obtained by lateral fusion of the characters along with their constituent characters, their occurrence in middle zone and their overall contribution. A total of 655(unique) such recognizable units corresponding to category-1 and category-2 have been identified.

**Table 2.** Some recognizable units obtained by lateral fusion of the characters.

| Horizontally fused consonants | Character Combination | % Occurrence with in all middle zone characters | Overall % Occurrence |
|---|---|---|---|
| प्र | प ् र | 0.3766 | 0.2900 |
| त्र | त ् र | 0.1606 | 0.1237 |
| स्त | स ् त | 0.1601 | 0.1233 |
| क्ष | क ् ष | 0.1332 | 0.1026 |
| न्ह | न ् ह | 0.1198 | 0.0923 |
| रु | र ु | 0.1121 | 0.0863 |
| क्य | क ् य | 0.1052 | 0.0810 |
| न्द | न ् द | 0.0949 | 0.0731 |

The table shown below (Table 3.) provides the overall contribution of these recognizable units if we select a definite number of these units. From this table (Table 3.) it is evident that if we select first 200 such units they will contribute to 2.9040%, and they all will contribute to 2.9349% toward overall recognition.

**Table 3.** Overall contribution of laterally fused recognizable units.

| Horizontally fused consonants selected out of 655 | % Contribution toward middle zone | Overall % contribution |
|---|---|---|
| 10 | 1.4286 | 1.1003 |
| 20 | 2.0878 | 1.6080 |
| 50 | 3.0632 | 2.3593 |
| 100 | 3.5570 | 2.7396 |
| 150 | 3.7094 | 2.8570 |
| 200 | 3.7705 | 2.9040 |
| 250 | 3.7915 | 2.9202 |
| 350 | 3.8043 | 2.9301 |
| 400 | 3.8066 | 2.9318 |
| 450 | 3.8079 | 2.9329 |
| 500 | 3.8089 | 2.9336 |
| 655 | 3.8106 | 2.9349 |

Similarly 209 (unique) and total of 317980 recognizable units corresponding to category-3 and category-4 have been identified. The table (Table 4.) depicts first few such units along with the frequency of their occurrence in middle zone and their overall occurrence**.**

**Table 4.** Overall occurrence of consonants with descender/vertically overlapped consonants.

| Consonants with descender/vertically overlapped consonants | % Occurrence with in all middle zone characters | Overall % occurrence |
|---|---|---|
| कु | 0.2879 | 0.2217 |
| ह्न | 0.2744 | 0.2113 |
| मु | 0.2734 | 0.2105 |
| सु | 0.2301 | 0.1772 |
| गु | 0.1380 | 0.1063 |
| पु | 0.1142 | 0.0879 |
| पू | 0.1141 | 0.0879 |

The following data provides the overall contribution of these recognizable units if we select a fix number of these units. From this data it is clear that if we select all such units they will contribute to 2.6206% toward overall recognition.

**Table 5.** Percentage contribution of consonants with descender/vertically overlapped consonants selected out of 209 recognizable units.

| Consonants with descender/vertically overlapped consonants out of 209 | % Contribution toward middle zone | Overall % contribution |
|---|---|---|
| 10 | 1.6988 | 1.3084 |
| 20 | 2.3778 | 1.8314 |
| 50 | 3.1215 | 2.4042 |
| 100 | 3.3695 | 2.5952 |
| 150 | 3.4000 | 2.6187 |
| 209 | 3.4025 | 2.6206 |

The total number of recognizable units counts to 942. These units are so large in number that it is very difficult to handle all these as separate classes, therefore we optimize the class count by considering all possible recognizable units and then evaluating their overall contribution. The table (Table 6.) depicts first few recognizable units and their contribution with in all 942 classes.

**Table 6.** Percentage occurrence of few recognizable units out of 942 recognizable units.

| S.No. | Recognizable unit | Overall % Occurrence |
|---|---|---|
| 1. | ◌ा | 22.9577 |
| 2. | ◌े | 8.2331 |
| 3. | क | 5.1207 |
| 4. | र | 4.9624 |
| 5. | ◌ी | 4.1946 |
| 6. | न | 3.7131 |
| 7. | ◌ि | 3.5408 |
| 8. | ह | 3.3730 |
| 9. | ◌ं | 3.3140 |
| 10. | म | 3.1325 |

From data given below (Table 7.) we find that if we chose first 100 recognizable units they will contribute to 98.0898%. Similarly the selection of first 600 units will result in the contribution of 99.9951%.

**Table 7.** Percentage contribution of recognizable units.

| Recognizable units | % contribution |
|---|---|
| 20 | 82.0185 |
| 30 | 90.1112 |
| 40 | 93.4336 |
| 50 | 95.0826 |
| 70 | 96.6985 |
| 100 | 98.0898 |
| 150 | 99.1658 |
| 200 | 99.5661 |
| 300 | 99.8817 |
| 400 | 99.9672 |
| 500 | 99.9883 |
| 600 | 99.9951 |
| 700 | 99.9976 |
| 800 | 99.9989 |
| 942 | 100.0000 |

## Conclusion

In Devanagari script each zone contributes to the different recognizable units. From character recognition point of view there are 78 distinct recognizable units in Devnagari script. But in Devnagari script the characters may fuse (Fig.1-4) resulting in compound characters. These characters may be very difficult to separate, and hence contribute to segmentation errors. So to avoid such errors we shall consider all such compound characters as separate recognizable unit. We have identified 864 such unique units. The overall recognizable units are obtained by adding basic 78 recognizable units and the 864 (Table 2, 4.) compound recognizable units, which result in a total of 942 recognizable units. As it would be very difficult to handle such a large number of classes, so the contribution of each class is evaluated to find how many classes would be sufficient to get desired recognition rate. From above table it is clear that if we consider first 100 classes they will contribute to 98.0898%, similarly first 150 classes will contribute to 99.1658%. It has also been found that the single recognizable unit '◌ा' contributes 22.9577% toward the overall accuracy.

## References


1. Sinha, R.M.K., Mahabala, H.N.: Machine recognition of Devnagari script. In: IEEE Transactions on Systems, Man and Cybernetics, vol. SMC-9, pp. 435--441 (1979)
2. Sinha, R.M.K., Bansal, V.: On Devnagari Document Processing. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp. 1621--1626 (1995)
3. Chaudhuri, B.B., Pal, U.: An ocr system to read two Indian language scripts: Bangla and Devnagari (Hindi). In: Proceedings of the 4th International Conference on Document Analysis and Recognition, vol. 2, pp. 1011--1015, Germany (1997)
4. Bansal, V., Sinha, R.M.K.: Integrating Knowledge Sources in Devnagari Text Recognition. In: IEEE Transactions on Systems, Man and Cybernetics-part A: Systems and Humans vol. 30, pp. 500--505 (2000)
5. Bansal, V., Sinha, R.M.K.: Segmentation of touching and fused Devnagari characters. In: Pattern Recognition, vol. 35, pp. 875--893 (2002)
6. Kompalli, S., Nayak, S., Setlur, S.: Challenges in OCR of Devnagari Documents. In: Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 327--331 (2005)