An Analysis of Irregularities in Devanagari Script Writing – A Machine Recognition Perspective

Satish Kumar
Department of Computer Science and Applications,
Panjab University Regional Centre,
Muktsar, Punjab, India

***Abstract*— This paper deals with the theoretical analysis of Devanagari script according to machine recognition perspective. An individual having good knowledge of the script of a language can easily read some words written on a paper pertaining to that script, though these are expressed in very bad manner, on the basis of his/her mental dictionary. Such words may not be easy to read through a machine as there may be various irregularities caused in expressing these words which are not easy to handle by a machine. A word in Devanagari script can be divided into three vertical regions i.e. the upper, the lower and the middle. The irregularities caused can belong to the upper, the lower and the middle region in a hand-printed word. Some common irregularities in writing Devanagari script are reported here. The various problems likely to be faced in machine recognition process pertaining to the words of Devanagari script are also detailed. Each individual writes uniquely. The writing of a writer may be very good or very bad. Some awkward styles of writing, due to which a lot of difficulties are faced in machine recognition process, are also discussed in detail.**

Keywords- Devanagari; printed; irregularities; machine-printed; intelligent character recognition(ICR).

I. INTRODUCTION

An ICR works in various stages such as scanning, pre-processing, feature extraction, classification and post-processing. Segmentation is one among the various pre-processing steps performed on an image before feature extraction and classification is carried out. The variability caused in script writing is so high that some times it becomes difficult even to read a hand-printed material by a well knower of the language and so reading the same script through a machine can not be expected.

Hindi is widely used language in south Asian region which is written in Devanagari script. Hindi is also official language of India. The words in Devanagari script are not written in cursive manner but there are some alphabets which some individuals write in cursive manner. Though it is non-cursive in writing but there are some irregularities which some writers commit in writing this script that poses some difficulties to read Devanagari script through a machine. Some such irregularities have been described in this paper.

Some papers dealing with the recognition of Devanagari script prior to 1990 have been referenced as:[2,5,6,7]. Two major work dealing with the recognition of machine-printed script of Devanagari after 1990 have been referenced as:[1,4]. Some papers are available on Devanagari off-line hand-printed character or word recognition in literature. Prominent among these are: Satish [13-14], Pal et al[15], Sharma et al[16] and Deshpande et al [17]'s works for isolated character recognition and Shaw et al[18] and Parui et al[19]'s works on Devanagari hand-printed isolated word recognition. Apart from this some papers dealing with the segmentation of Devanagari script have been referenced as: [11, 12] and some papers dealing with the recognition of on-line hand-printed Devanagari script have been referenced as: [8, 9].

The research work is financially supported by UGC, India.

II. A BRIEF SCRIPT REVIEW

In Devanagari script, a word can be vertically divided into three regions like Roman: the upper region, the middle region and the lower region. The upper region is occupied by the vowel symbols or a part of vowel symbols; the lower region is also occupied by the vowel symbols while the middle region is occupied by the consonants, pure consonants, compound characters, vowels, vowel symbols or a part of vowel symbols. A machine-printed and hand-printed Devanagari word demonstrating all the three vertical regions are given in "Fig. 1" and "Fig. 2" respectively.

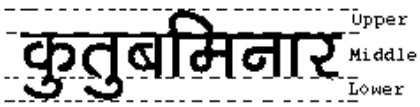


Figure 1. A machine-printed Devanagari word divided into three vertical regions.

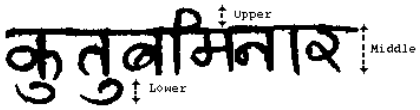


Figure 2. A Hand-printed Devanagari word divided into three vertical regions.

The size of all the three regions is not same in case of machine-printed and depends upon font to font. The lower and upper regions are some what smaller in size as compared to middle one. Keeping in view the three regions it is not difficult to segment a machine-printed Devanagari word. In case of hand-printed the size of various regions is not expressible i.e. the size of a region can be bigger or smaller and depends upon

a writer's writing style. Some irregularities which are caused by the writers at the writing moments are discussed in next Section. Due to the irregularities caused in writing, a lot of efforts are required to read the given script through a machine. Some times, the written material is so awkward that it becomes difficult to read even visually. The solution to recognize such writing is some what difficult but not impossible. To improve the recognition accuracy of an ICR, either we have to teach the masses to avoid bad writing or we have to design a robust algorithm to deal with such complex issues.

III. SCRIPTING DISCREPANCIES

There are various irregularities which are performed by a writer at the writing moments that makes word level recognition of hand-printed Devanagari script a lot of difficult. The cause of irregularities may be hasty writing, habitual of bad writing, lack of interest in writing, lack of concentration, lack of script knowledge, lack of proper environment, bad instrument and bad stationary. Even if last four factors are favourable the irregularities in writing can not be ignored. The various issues pertaining to such discrepancies are explained region wise as:

A. Upper Region

The upper region in a Devanagari word is occupied by the vowel symbols or a part of the vowel symbols. The various vowel symbols or a part of symbols used in this region are given in Table 1. Some common irregularities are as:

TABLE 1: VOWEL SYMBOLS OR A PART OF VOWEL SYMBOLS IN UPPER AND LOWER REGION.

Region	Symbols/a part of symbols
Upper	◌ ◌ा◌ि ◌ी◌ु ◌ू◌े ◌ै◌ो ◌ौ
Lower	◌ृ ◌ॄ ◌ॢ ◌ॣ

1) Abnormal size of a vowel symbol: Some individuals write a vowel symbol or a part of a vowel symbol very small in size where as some individuals write the same symbol very large in size in contrary to its actual size. Actually the size of a vowel symbol or a part of a vowel symbol in a word should be balanced. Some examples of unbalance and balanced vowels in upper region are shown in Fig 3. In all the four words, mentioned in "Fig. 3", the size of the vowel symbols is so small that it is difficult to perceive and conclude the identity of the symbols in isolation. Such irregular writing not only causes difficulties in recognizing a word through a machine but also by an individual.

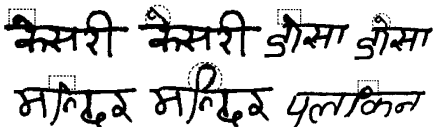
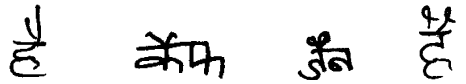


Figure 3. Unbalanced vowel symbols encircled inside dotted rectangles and balanced vowel symbols encircled inside dotted circles.

2) Awkward representation of a vowel symbol: Some individuals write a part of a vowel symbol 'ा' above head-line as given in "Fig. 4(a)". Writing this symbol as a straight stroke do not pose any problem in recognition but two such strokes above head-line are some times written in a very lumbering way as given in "Fig. 4(b-c)". Some writers also express the same vowel symbols, as given in "Fig. 4(d)", with extra branches. The awkward ways of expressing the vowel symbols, as given in "Fig. (b-d)" , above head-line are also irregularities in writing and may create a lot of problems in recognition process.

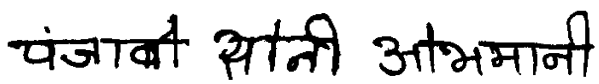


a) b) c) d)

Figure 4. Some good and bad vowel symbols above head-line.

3) Incomplete and inaccurate representation of a vowel symbol: In some cases the upper part of the vowel symbols 'ि' and 'ी' is written in an ugly way. Though a well knower of the script can easily interpret it but it may pose a lot of difficulties in machine recognition. Some such ugly represented upper vowel symbols are given in "Fig. 5".

Figure 5. Some words with ugly written upper vowel symbol.



4) Merging vowel symbols with head-line: In some cases it has been observed that a vowel symbol is completely amalgamated in head-line. If this fusion is partial then it may not pose any problem in segmentation and recognition process but in cases where this fusion is large which completely destroy the structure of a vowel symbol may cause a lot of problems in recognition process. In "Fig. 6", in case of word भिंडी, a vowel symbol 'Bindi' has been touched with head-line. The recognition process may consider it as a blot along head-line and can ignore it being a part of head-line. There should be a small gap between a 'Bindi' and head-line. "Fig. 6" shows some such irregularities.

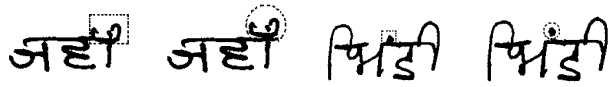


Figure 6. Completely merged vowel symbols with head-line encircled in dotted rectangle and properly placed vowel symbols adjacent to head-line encircled in dotted circle.

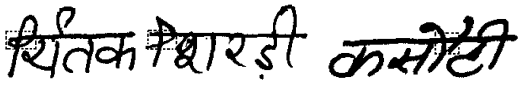


Figure 7. Some words with wrong attachment of vowel symbols.

5) Intruding upper vowel symbols with middle region: Some writers wrongly originate the vowel symbols in upper region. In some words these are not originated from the head-line rather these are originated from middle region. Consequently, a portion of these remain in middle region that also poses a lot of problems in segmentation and recognition process. Some such words with wrong attachment of the vowel symbols are reported in "Fig. 7".

B. Lower Region

The lower region in a Devanagari word is occupied by the vowel symbols only. Some common irregularities observed in this region are as:

1) Intruding a lower vowel symbol with middle region: A majority of writers do not write the lower region vowel symbols in lower region but intrudes the same to a large extant in middle region. In some cases such encroachments are so large that the lower vowel symbols are completely inside middle region as given in "Fig. 8". In such cases it becomes difficult to judge the existence of a vowel symbol attached to the lower part of a consonant.

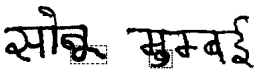


Figure 8. Word images where lower vowel symbols completely inside middle region.

2) Improper attachment of a lower vowel symbol: In some cases the lower vowel symbols are not written at their proper place. The cause may be habitual writing or un-controlled writing. In such situation it becomes more difficult to identify a lower vowel symbol. Some examples are given in "Fig. 9". In all these three cases the writers have attached a lower vowel to a consonant by ignoring scripting rules. In such cases too it becomes difficult to judge the existence of a vowel symbol attached to the lower part of a consonant.

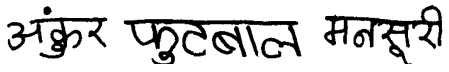


Figure 9. Word images where lower vowels are attached to consonants ignoring scripting rules.

3) Chomping a consonant as a result of wrong attachment of a lower vowel symbol: In some cases a lower vowel symbol completely overlaps a consonant to which it is to be connected and destroys its shape that is very difficult to recognize. The examples of such words are given in "Fig. 10". In first word it appears as a lower vowel symbol 'ा' is attached to 'ए' but actually it is attached to 'य'. So it has completely destroyed the structure of 'य'. Similarly in second word, a vowel symbol 'ा' has destroyed character 'द' from lower side and consequently it is appearing as 'ट'.

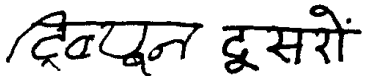


Figure 10. Word images where lower vowels chomped the consonants.

4) Writing a vowel symbol in Isolation: Some persons have habit of not attaching a lower vowel symbol to a consonant rather writes it quite away. If there is not any line below a line in which this word is written then there is no problem. However, if there is a line below a given word then some times it becomes difficult to draw conclusion whether this isolated vowel symbol is attached to a word written in lower line as an upper vowel symbol or it is attached to a word written in upper line as a lower vowel symbol. The situation is given "Fig. 11".

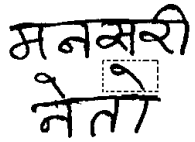


Figure 11. Two words with a confusing vowel symbol marked in dotted rectangle.

5) Abnormally expressed vowel symbols: One more important observation about the lower vowel symbols is that some persons write a lower vowel symbol very large as compared to its actual size relative to a word as given in "Fig. 12". Such irregularities are natural as some times a writer losses control at the time of writing but it may cause problem in machine recognition process.

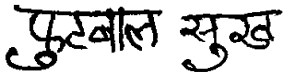


Figure 12. Some words with an enormous lower vowel symbol.

C. Head-line

The characters in a Devanagari word are connected to each other through a head-line. At writing time, generally, a head-line is inserted after writing all the characters or character strokes in a word. Consequently, the whole line in a word is written as a single stroke. In all above mentioned examples given in "Fig. (2-10)", the head-line has been inserted as a single stroke. But some persons have habit of step-wise writing where head-line is inserted after writing each character in a word. Consequently, a head-line is zigzagged. Some people have also a habit of inserting an incomplete head-line or a broken head-line. The examples of such words are given in "Fig. 13". The head-line is most important part of a Devanagari word that can be exploited to know and recognize it effectively. If head-line in a word is not smooth then it becomes difficult to locate the various components or symbols in a word and thus contributes to drop in recognition rate.

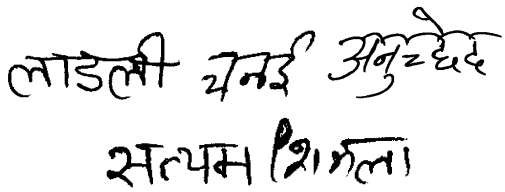


Figure 13. Devanagari words with incomplete, broken and zigzagged head-lines.

D. Middle Region

The middle region may contain the consonants, the vowels or the vowel symbols. In other words one can say that middle part may contain all vowel symbols of Devanagari except the vowel symbols used in upper and lower region in addition to consonants, vowels etc. There are various irregularities which are committed by a writer while writing the script in this region. Some such irregularities are explained as:

1) Wrong insertion of head-line: This irregularity can be considered belonging to both head-line and middle region. This kind of irregularity is not due to the habitual writing and may be caused in hasty writing. The head-line always encroach a lot of area of middle region and consequently all the characters of this region are destroyed a lot. This causes a lot of difficulties in segmenting and recognizing the given characters in a word which in some cases are 10-30% destroyed from top. Though such words can be easily recognized visually but may pose problem in machine recognition process. Some completely destroyed words by wrongly inserting head-line are given in "Fig. 14".

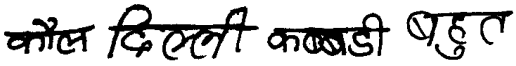


Figure 14. Some completely destroyed characters in a word due to wrong insertion of head-line.

2) *Incomplete character writing*: Some writers also write incomplete characters in middle region. It may be easy to judge the identity of a character on the basis of the other consonants, vowels or vowel symbols, etc in a given word as a person's brain itself is a big dictionary. In case of machine recognition the judgment about the identity of an ambiguous character can be made only in post-processing stage based on a dictionary of words. Some words with incomplete characters and their exact character structures are given in "Fig. 15".

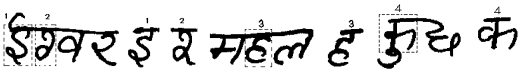


Figure 15. Some words with incomplete characters and their exact character structure.

3) Touching characters with head-line due to wrong attachment of a character with it: Some persons touch some characters in a word with head-line and are completely destroyed where as remaining characters are in good condition. Such destroyed characters are difficult to recognize. Actually, the main cause of touching some characters with head-line is that there is a very small gap between head-line and a stroke in structures of some characters. In Devanagari the various such characters are इ क ड न ब व , etc in which the gap between head-line and their upper stroke is very small. In such cases some times a writer fails to keep required gap due to fluent writing. Some characters touching with head-line are given in "Fig. 16".

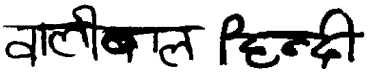


Figure 16. Some words with some damaged character due to head-line touch along with some reasonable characters

4) *Narrow writing*: Some writers have habit of writing consecutive characters very narrow and they do not keep required gap between two characters. The result is that a lot of characters in a word are touching. In some cases the touching is partial where as in other cases the touching is very large and it completely destroys the structure of both the characters. Some words written without/negligible gap between two characters are given in "Fig. 17".

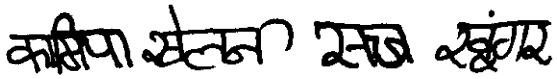


Figure 17. Some congested words.

5) *Overwriting*: Some times in writing a stroke used to express a character may be faded or broken. To ensure the type of stroke or symbol in a character, a writer over-writes the given character or a part of a character. This behavior widens the stroke width which is always different from the various strokes presented in rest part of a word or a character. Some times even extra branches or extra loops are formed which changes the structure of a character a lot and contributes to drop of recognition rate of an ICR. Some such characters are demonstrated in "Fig. 18".

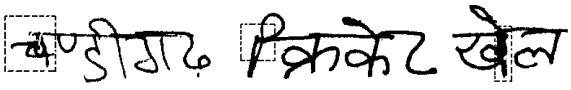


Figure 18. Some words where some characters destroyed due to overwriting.

IV. Some Awkward Styles

Each individual has his own style of writing. About writing one can not question about how and why a person writes this way or that way. As far as about the style of one's writing is concerned, a person acquires it from lower classes. But if one has determination to change it, he/she can do easily. As per

my perception, it is related to one's mental ability. Apart from the above mentioned irregularities, some awkward styles of writing the various characters in a word are as:

A. *Undeveloped Character Writing*

Some persons have habit of writing undeveloped or partial characters. They write a lot of characters in a word in incomplete form. Also there are some characters which are partially written while a writer writes in flow. In case of machine recognition the judgment about the identity of such characters can be only made in post-processing stage based on a dictionary of words. Such characters, if written, in full developed manner will be very helpful to enhance the recognition rate of an ICR. Some such examples are given in "Fig. 19".

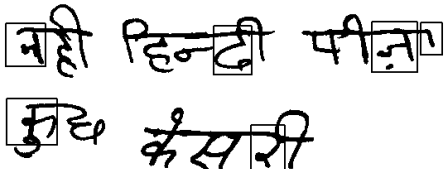


Figure 19. Some partially or incompletely written words.

B. *Unusual Character Writing*

Some writers are habitual of writing very enormous characters. They write in a very strange manner. The structure of a hand-printed character is quite different from the regular structure of a character. Such writing also causes difficulties in machine recognition process. Some such character images are given in "Fig. 20".

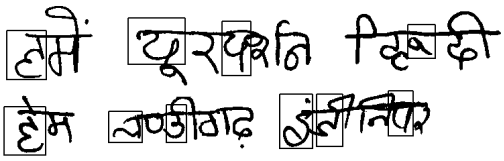


Figure 20. Some words having enormous characters those are difficult to recognize even visually.

C. *Extra Stroke Formation*

Some writers create extra branches in writing which are usually not required. The extra strokes created by them are not toward decorating a character/word to convert it to stylish rather as a result of habitual writing or hasty writing. Some times extra branches formed are so large in number and size that these make the structure of a character/word quite gauche. Some examples of such writing are given in "Fig. 21".

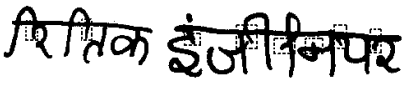


Figure 21. Two words with unwanted strokes shown in dotted rectangle region.

D. *Touching Strokes*

When a writer writes, he/she may not be able to keep control on his/her hand. Consequently, a lot of characters/ character symbols touch with each other. Actually, this touch is of two types: inter-character and intra-character.

1) Inter-character: In such kinds of touching, the two adjoining characters touch with each. The fusion at touching point may be partial or large. In such situation it becomes difficult to segment and recognize the fused characters as it becomes difficult to identify the exact boundary between these characters. Some examples are given in "Fig. 22".

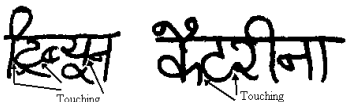


Figure 22. Examples of inter-connected characters in a word.

2) Intra-Character: In such kinds of touching, the two or more strokes in a character touch with each other. The fusion at touching point may be partial or large as in case of inter-character touch. Some times the touch is so large that it completely destroys the given character. In such cases it becomes difficult to judge the identity of the given character in isolation. In machine-recognition process it becomes difficult to recognize such characters. Some such examples are given in "Fig. 23". Such characters some times are difficult to identify visually in isolation.

Figure 23. Examples of inter-connected characters in a word.

V. Discussion and Conclusion

There is a lot of difference between hand-printed and machine-printed. In hand-printed writing a lot of irregularities are committed by the writers. These irregularities drop the recognition rate of an ICR a lot. Though, there are some issues which are beyond the control of a writer, as a writer some times losses control at the writing moments, even then the drop in recognition rate of an ICR can be improved by teaching the masses about avoiding bad writing and by designing a robust algorithm to deal with such complex issues.

References

[1] V, Bansal, "Integrating knowledge sources in Devanagari text recognition system", Ph. D thesis , 1996.

[2] I. K. Sethi and B. Chatterjee, "Machine recognition of constrained hand-printed Devanagari numerals", J. institute of electronic telecommunication engineering, vol. 22, pp. 532–535 , 1976.

[3] R. Bajaj , L. Dey and S. Chaudhury, “Devanagari numeral recognition by combining decision of multiple connectionist classifiers”, Sadhna, vol. 27, Part 1, pp. 59-72 , 2002.

[4] U. Pal and B. B. Chaudhuri, “Printed Devanagari script OCR system”, Vivek, vol. 10, pp. 12-24, 1997.

[5] R. M. K. Sinha and H.N. Mahabala "Machine recognition of Devanagari script", IEEE transactions on systems, man, and cybernetics, vol. SHC-9, no.8, pp. 435- 441 , 1979.

[6] K. Jayanthit, A. Suzukit , H. Kanait , Y. Kawasoej, M. Kimurat and K. Kido, “Devanagari character recognition using structure analysis”, Fourth IEEE region 10th international conference, Bombay, India, pp. 363-366, 1989.

[7] I. K. Sethi and B. Chatterjee, Machine recognition of constrained hand-printed Devanagari, Pattern Recognition, vol. 9, no. 2, pp. 69-75 , 1977.

[8] S. D. Connel, R.M.K. Sinha and A. K. Jain, “Recognition of unconstrained on-Line Devanagari characters”, Proceedings of the international conference on pattern recognition, Barcelona, Spain, vol. 2, pp. 368-371, 2000.

[9] N. Joshi, G. Sita and A.G. Ramakrishan, “Machine recognition of online handwritten Devanagari characters”, Proceedings of the eight international conference on document analysis and recognition, vol. 2, pp. 1156- 1160, 2005.

[10] V. Bansal and R.M.K. Sinha, “Partitioning and searching of dictionary for correction of optically read Devanagari character strings”, International journal on document analysis and recognition, pp. 269-280, 2002.

[11] U. Garain and B. B. Chaudhuri, “Segmentation of touching characters in printed Devanagari and Bangla scripts using fuzzy multifactorial analysis”, IEEE transactions on systems, man, cybernatics- part C: applications and reviews, vol. 32, no. 4, 2002.

[12] V. Bansal and R. M. K. Sinha, “Segmentation of touching and fused Devanagari characters”, Pattern Recognition, vol. 32, pp. 875-893 , 2002.

[13] Satish Kumar, “A three tier scheme for Devanagari hand-printed character recognition”, Proceedings of 8th international conference on computer information systems and industrial management applications, Coimbatore, India , 2009.

[14] Satish Kumar, “Performance comparison of features on Devanagari hand-printed dataset”, International journal of recent trends in engineering, vol. 1, no. 2, pp. 33-37, 2009.

[15] U. Pal, N. Sharma, T. Wakabayashi, F. Kimura, "Off-line handwritten character recognition of Devanagari script", Proceedings of 9th international conference on document analysis and recognition , vol. 1, pp. 496-500, 2007.

[16] N. Sharma, U. Pal et al ,”Recognition of off-line handwritten Devanagari characters using quadratic classifier”, ICVGIP 2006, LNCS 4338, pp. 805-816, 2006.

[17] P. S. Deshpande, L. Malik and S. Arora, “Fine classification & recogntiton of handwritten Devanagari characters with regular expression & minimum edit distance method”, Journal of computers, vol. 3, no. 5, pp. 11-17, May 2008.

[18] B. Shaw, S. K. Parui and M. Shridhar, “Off-line handwritten Devanagari word recognition: A segmentation based approach”, IEEE , 2008.

[19] S. K. Parui and B. Shaw, “Off-line Devanagari handwritten word recognition: An HMM based approach”, Proceedings of PReMI-2007(Springer), LNCS-4815, pp. 528-535, 2007.

AUTHORS PROFILE

Dr. Satish Kumar is a faculty member of Punjab University, Chandigarh (India), currently posted at Punjab University Regional Centre, Muktsar , Punjab, (India) and has about ten years experience of teaching Post-graduate classes and actively engaged in research activities. His areas of interest are Image processing, Pattern recognition and artificial intelligence.