Segmentation of Isolated Handwritten Marathi Words

C. H. Patil
Research Scholar, Department of Computer Science, Yashwantrao Mohite College, Bharati Vidyapeeth Deemed University, Pune

S. M. Mali
Department of Computer Science, MACS College,Pune
Savitribai Phule University of Pune , Pune

ABSTRACT
In this paper, database for isolated handwritten Marathi simple words (not contains compound character known as '*Jodakshare*') was developed. Commonly used 50 words are chosen and total 20210 handwritten Marathi word samples database was developed. Generalized segmentation methodology is proposed here which is applicable to any simple, handwritten Marathi word containing any number of characters. The segmentation algorithm first detects header cap ('*Shirorekha*') that separates top modifiers and core area of the word. Statistical information and vertical projection is used for further segmentation process. The segmentation algorithm proposed here is applicable to any Marathi plain word having any number of characters also can equally applicable to many other languages like Hindi, Sanskrit, Nepali and Konkani which are similar in structure. Using proposed algorithm maximum 95.31 percent segmentation is achieved for '*Daar*' word and average segmentation achieved is 82.17 percent.

Keywords
Handwritten Marathi word, segmentation, image, vertical projection, database development

1. INTRODUCTION
Marathi is well known language spoken by people of Maharashtra. There are around 100 million speakers of Marathi language which is fourth largest number of native speakers in India. Marathi belongs to Indo-Aryan group of languages. Sanskrit is the origin of all Indo-Aryan languages. Balbodh script is currently used for Marathi language. Balbodh is originated from Devanagari script. Marathi also has influence of other languages like Sanskrit, Kannad, and Telugu. Lots of words are entered into Marathi from Persian, Turkish and Arabic. Also Portuguese and British influenced Marathi through their words.

There are two approaches for handwritten text recognition. First approach is holistic approach which is more useful if words are limited. In this approach directly features are extracted from word samples and classified. But Marathi language consists of unlimited words so using first approach is not suitable. Second approach is segmentation based approach in which handwritten Marathi words are divided into isolated indivisible characters, and further these indivisible characters are used for classification process.

The problem of segmentation and difficulties in segmentation are well studied and reported in the literature. Old typewritten Gujarati document segmentation problem is taken by apurva desai [1] by using vertical projection method achieved 65% result. Bikash Shaw [2], Brijmohan Singh et la [5] also reported projection method for handwritten word recognition. Dipankar Das et al [6] reported best cut method for touching Bangala numerals and achieved 89.7% result. Naresh Kumar Garg et al [9] has considered problem of segmentation of Hindi text and reported 79.12% result by using vertical projection method. Morphological approach for segmentation of handwritten Devanagari text is reported by Sandip N.Kamble [14] and achieved 52% result. Suryaprakash Kompalli et al [15] reported graph based segmentation approach and achieved 85% result.

To recognize handwritten Marathi text segmentation is very important phase which divides text into line, lines into words and then words into characters. In this part of work we have considered isolated handwritten Marathi words as input and segmenting them to isolated indivisible characters.

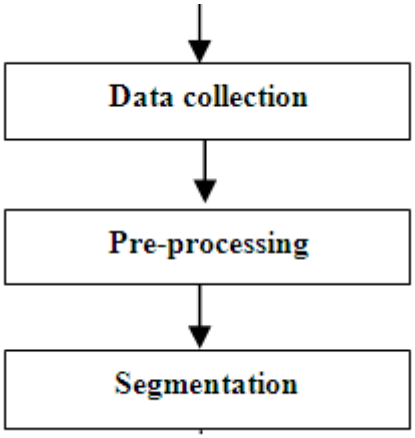


Fig 1: Steps for handwritten Marathi word segmentation

Phases for handwritten Marathi word segmentation are shown in the Fig. 1. In this paper Section 2 focusing on development of database for isolated handwritten Marathi word segmentation, Section 3 discuss pre processing techniques applied, section 4 discuss segmentation phase and section 5 the findings.

2. DATA COLLECTION
To recognize handwritten Marathi word first attempt is to develop database for handwritten words. For handwriting recognition system CEDAR, NIST and CENPARMI databases are usually get used. All these databases are for isolated characters, numerals or words for English language. But benchmark database for handwritten Marathi word is not exist [2]. Sufficient work is reported for Devanagari characters and Marathi characters. Handwritten Devanagari word database reported by Bikash Shaw et al. [2, 3] for 100 words collected samples 39700 from 436 writers. Handwritten Devanagari word databse for 30 classes 28,500 samples are collected by Brijmohan Singh et al [4]. Handwritten Marathi word database for 10 classes of Devanagari numerals in word form from one to ten for each class 200 samples total 2000 samples database were reported by Laurent Guichard et al [8]. Handwritten Marathi 114 words of legal amounts database is reported by R. Jayadevan et al [12]. Handwritten Hindi script 100 blocks are taken as database by G. G. Rajput et al [7]. Handwritten Hindi 1380

words of 200 lines are collected by Naresh Kumar Garg et al [9-11]. 2,000 constrained and 2,000 unconstrained Devanagari words are collected by Rajiv Kumar et al[13]. Devanagari 100 words are collected by Sandip N.Kamble et al[14]. Database of handwritten 3000 Marathi word was reported by Vijaya Rahul Pawar et al [18].

Literature review indicates that benchmark database for handwritten Marathi word is not available to carry experiments. Since a benchmark database is not available, first attempt was to develop sufficient database for handwritten Marathi words. Commonly used 50 Marathi words are selected and three A4 size sheets were designed for data collection and distributed to 50 writers which include students, teachers, and clerks. Every writer has to write a word in square provided and no other constraint was imposed on writers.

The data sheets were scanned using a flat bed scanner at a resolution of 1200 dpi and stored as gray scale images. From the scanned gray scale image, word images were cropped manually and stored in respective class folders. For each word approximately 500 samples were stored in respective class folders. Finally database of 20210 handwritten Marathi word images are ready to carry experiments. Fig. 2 shows gray scale handwritten Marathi word images cropped from the scanned image of a datasheet.

3. PREPROCESSING

Pre-processing commonly involves in normalizing the intensity of the individual particles images by removing reflections and masking portions of the images. Pre-processing enhances word images for correctly segmenting words into isolated characters and modifiers symbols.

The raw input for the digitizer typically contains noise due to erratic hand movements and inaccuracies in digitization of the actual input. In order to reduce the blurring of character edges and suppress noise, the median filter is used.

Fig 2: Dataset of handwritten Marathi words

In median filtering, the idea is to replace the current point in the image by the median of the brightness in its neighborhood. A 3×3 square neighborhood is used to remove noise from the gray scale images.

Image binarization is performed on input image. Histogram-shape based image thresholding suggested by Otsu's is used for converting gray scale image to binary image.

The algorithm assumes that the image contains two classes of pixels (foreground and background) prior to thresholding and it calculates the optimum threshold separating those two classes so that their combined spread (intra-class variance) is minimal.

The binarized image is mapped onto a standard plane (with predefined size) so as to give a representation of fixed dimensionality for classification. The goal of image normalization is to reduce the inter-class variation of the shapes of the words in order to facilitate segmentation process and improve their segmentation accuracy and ultimately recognition accuracy. Linear normalization method is used to standardize the word images. The standard plane is considered as a square of size 60 x 60. The width and height ratio of the word image is not disturbed due to normalization.

The goal of image thinning is to remove pixels so that an object without holes shrinks to a minimally connected stroke, and an object with holes shrinks to a ring halfway between the hold and outer boundary.

4. SEGMENTATION

Segmentation based approach of recognition of handwritten Marathi words is dependent on the success of segmentation phase. Segmentation is very difficult and challenging task in handwritten text recognition because of various reasons described as follows.

4.1 Difficulties in Segmentation

4.1.1 Shirorekha

Marathi has most prominent characteristics in every word called header cap known as '*Shirorekha*' which is written from left to right on the top of characters in the words. Sometimes writers don't write '*Shirorekha'* or is broken or slanted or is mixed with characters. Detecting location of Shirorekha is important part for segmentation process. If location of Shirorekha is not detected correctly segmentation of word fails due to this classification process fails.

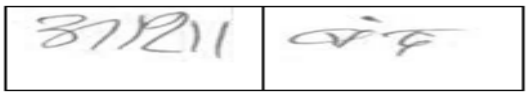


Fig 3: Words where no '*shirorekha'* written

4.1.2 Touching characters

Due to irregular handwritings it may happen that characters are touched to each other or connected to the modifiers of other characters. Touching characters creates problem during segmentation of words into isolated characters.

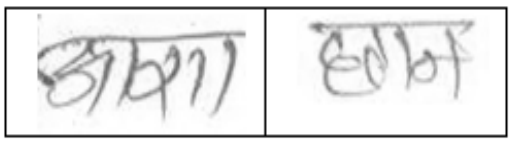


Fig 4: words having touching characters

4.1.3 Slanting characters

Due to different handwriting styles or style of keeping paper it may happen characters in the words are slanted. Due to this detecting location of shirorekha is difficult which hampers segmentation process.

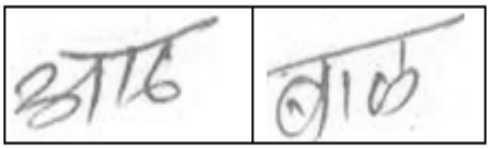


Fig 5: Words having slanted characters

4.1.4 Broken characters

Due to various problems like instrument for writing not working properly, writing style, poor quality paper or damaged papers can cause broken characters. If characters are broken in the word it may cause over segmentation of character. Over segmentation reduces the recognition accuracy of word recognition.

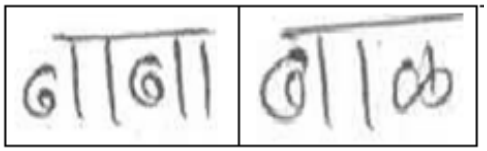


Fig 6: Words has broken characters

4.1.5 Overlapping characters

It may happen characters are overwritten due to improper writing or if writer is in hurry. Also modifiers are overwritten on characters. Due to overlapping characters segmentation phase fails to segment isolated characters and modifiers.

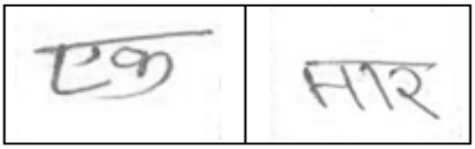


Fig 7: Words has overlapping characters

4.2 Segmentation Methodology

Proposed method of segmentation uses statistical information and vertical project to segment handwritten word into indivisible characters. Algorithm 1 uses header cap to divide top modifiers and core area of the word. After removal of header cap vertical projection is used to segment characters and modifiers present in the core area. Algorithm 2 segments top strip modifiers using vertical projection and assign top modifiers to respective characters in the core word.

Algorithm 1

Input: Handwritten Marathi Word Image

Output: Segmented isolated characters.

1. Read input handwritten Marathi word image.

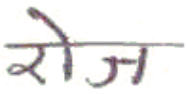


Fig 8: Sample handwritten Marathi word

2. Perform pre-processing on input image.

Fig 9: Preprocessed handwritten Marathi word

3. Calculate horizontal projection for the word image.

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 2 2 1 1 1 1 1 1 1 1 1 1 3 **45** 15
3 3 3 3 5 12 6 6 7 6 4 5 4 5 2 0
0 0 0 0 0 0 0 0 0 0 0 0 0

4. Find out the row number known as header line (*Shirorekha*) which contains maximum number of white pixel.

   Row number = 31

5. Convert all white pixels to black pixels of the header line identified in step 4.

Fig 10: word after removing *'Shirorekha'*

6. Divide word image into two parts. First part above '*Shirorekha'* cropped from the word image labeled as top_strip of that word contains top modifiers if any and second part labeled as core area of the word.

Fig 11: words top strip and core area

7. Calculate vertical projection for the core area of the word image.

0 **1** **3** **3** **3** **3** **2** **2** **2** **3** **3** **4** **7** **2** 0
0 0 0 0 0 0 **14** 0 0 0 0 0 0 **2** **1**
1 **1** **1** **1** **2** **2** **3** **5** **1** **1** **1** **1** **1** **1** **3**
4 **7** **3** 0 0 0 0 0 0 0 0 0 0 0 0

8. Find number of segments present in the core area of word by using vertical projection.

   Number of segments=3.

9. The number identified in step 8 is the number of isolated characters present in the image.
10. Repeat steps from 11 to 16 for the number of isolated characters present.
11. Skip all zeros.
12. Find out first column location contains nonzero value labeled as starting point.
13. Skip all nonzero numbers till zero.
14. Assign column location - 1 to the end point.
15. Crop the word image from starting point to end point column numbers.

Fig 12: Segmented indivisible characters

16. Assign remaining image to the core area of word image.
17. Segment top modifiers in top_strip and assign modifiers to respective segmented character in core area using Algorithm 2.

Algorithm 2

Input: Top strip image identified in step 6 Algorithm 1.

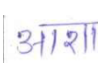
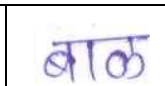
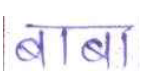
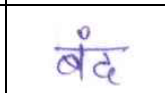
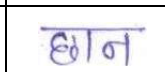
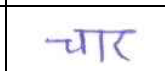
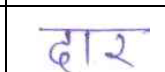
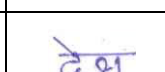
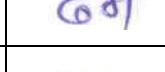
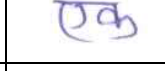
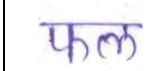
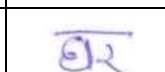
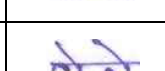
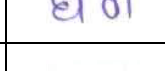
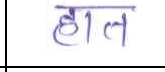
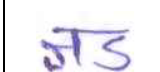
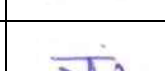
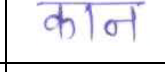
Output: Segmented top modifiers and assign to core area word segments.

1. Calculate vertical project for the top_strip.
2. Find number of segments present in top_strip.
3. Repeat step 4 to 6 for number of segments
4. Find end point of the segment.
5. If end point of segment is greater than starting point and less than ending point of any segment in core area word image assign top_strip segment to the segment of core area.
6. Otherwise if end point of segment is greater than core area segments end point and less than starting point of next segment then assign top_strip segment to first segment of core area.

5. RESULTS

Database of 20210 samples is used to carry out segmentation experiments. Table 1 shows segmentation results for every isolated handwritten Marathi word. Maximum segmentation result 95.31% is achieved for the word '*Daar'*. During testing it has been observed that due to '*shirorekha'*, touching characters, overlapping characters, broken characters and slanted characters reduces result of segmentation. Table 2 shows comparison of segmentation methodologies reported. As shown in Table 2 database of 20210 samples used in this work is quite large than databases used in literature. Proposed segmentation methodology is tested and achieved 82.17% average segmentation result.

Table 1. Segmentation result for handwritten Marathi word

SR. NO.	WORD	SAMPLES	RESULT IN PERCENTAGE
1		324	**79.01**
2		324	**90.43**
3		357	**84.31**
4		335	**70.45**
5		373	**92.49**
6		384	**80.47**
7		373	**93.83**
8		357	**87.68**
9		363	**82.92**
10		384	**95.31**
11		373	**88.47**
12		457	**94.09**
13		477	**77.78**
14		346	**80.06**
15		477	**91.61**
16		384	**85.16**
17		371	**85.71**
18		477	**81.34**
19		477	**87.00**
20		457	**88.84**
21		406	**77.34**
22		384	**63.28**
23		374	**58.56**
24		384	**78.13**

25		488	**88.93**
26		357	**77.87**
27		466	**88.84**
28		384	**69.27**
29		467	**85.65**
30		358	**65.64**
31		360	**86.11**
32		477	**76.52**
33		368	**73.64**
34		372	**83.06**
35		384	**89.58**
36		396	**67.17**
37		395	**82.28**
38		371	**80.32**
39		364	**77.75**
40		477	**86.37**
41		477	**87.63**
42		357	**70.87**
43		357	**68.07**
44		477	**91.40**
45		368	**81.52**
46		477	**84.91**
47		479	**88.73**
48		455	**87.69**
49		384	**86.72**
50		477	**87.84**
Total word Samples		**20210**	**82.17**

Table 2. Results reported in literature for segmentation

PN	Author	Language	Method	Result in Per	Samples
[1]	Apurva A. Desai	Gujrathi	Vertical Projection	65%	NA
[2]	Bikash Shaw et al	Devanagari	Morphology opening	NA	NA
[6]	Dipankar Das	Bangla	Best cut (touching char)	89.7%	NA
[9]	Naresh Kumar Garg	Hindi	Vertical Projection	79.12%	1380
[14]	Sandip N.Kamble et al	Devanagari	Morphology	52%	100
[15]	Suryaprakash Kompalli	Devanagari	Graph	85%	NA
PM	**C. H. Patil**	**Marathi**	**Statistical information and Vertical Projection**	**82.17%**	**20210**

*PN-Paper number

*PM- Proposed Method

*Per- Percentage

6. CONCLUSION

In this paper, database of 20210 samples of isolated handwritten Marathi commonly used 50 words was developed. Segmentation methodology used is based on the '*shirorekha*' detection which separates top modifiers and core area of the word. Segmentation methodology presented here uses statistical information and vertical projection. The algorithm proposed here applicable to any Marathi plain word having any number of characters also can equally applicable

to many other languages like Hindi, Sanskrit, Nepali and Konkani which are similar in structure.

7. ACKNOWLEDGMENTS

The author is grateful to Dr. M.S. Prasad, for their guidance, helpful discussion and encouragement during this work.

8. REFERENCES

[1] Apurva A. Desai. Segmentation Of Characters From Old Typewritten Documents Using Radon Transform. International Journal Of Computer Applications (0975 – 8887) Volume 37– No.9, January 2012.

[2] Bikash Shaw, Swapan Kr. Parui, Malayappan Shridhar. Offline Handwritten Devanagari Word Recognition: A Segmentation Based Approach. 978-1-4244-2175-6/08/$25.00 IEEE2008

[3] Bikash Shaw, Swapan Kr. Parui, Malayappan Shridhar. Offline Handwritten Devanagariword Recognition: A Holistic Approach Based On Directional Chain Code Feature And HMM. International Conference On Information Technology IEEE ICIT 978-0-7695-3513-5/08 2008

[4] Brijmohan Singh, Ankush Mittal, M.A. Ansari, Debashis Ghosh. Handwritten Devanagari Word Recognition: A Curvelet Transform Based Approach. International Journal On Computer Science And Engineering (IJCSE) Vol. 3 No. 4 ISSN : 0975-3397 40634

[5] Brijmohan Singh, Nitin Gupta, Rashi Tyagi, Ankush Mittal, Debashish Ghosh. Parallel Implementation Of Devanagari Text Line And Word Segmentation Approach On GPU. International Journal Of Computer Applications (0975 – 8887) Volume 24– No.9, June 2011

[6] Dipankar Das Rubaiyat Yasmin. Segmentation And Recognition Of Unconstrained Bangla Numeral. Asian Journal Of Information Technology 5(2)2006

[7] G. G. Rajput, Anita H. B.. Handwritten Script Recognition Using DCT And Wavelet Features At Block Level. IJCA Special Issue On "Recent Trends In Image Processing And Pattern Recognition" RTIPPR, 2010.

[8] Laurent Guichard Alejandro H. Toselli Bertrand Couasnon. Handwritten Word Verification By SVM-Based Hypotheses Re-Scoring And Multiple Thresholds Rejection. 12th International Conference On Frontiers In Handwriting Recognition 978-0-7695-4221-8/10 $26.00 IEEE 2010

[9] Naresh Kumar Garg, Lakhwinder Kaur, M. K. Jindal. Segmentation Of Handwritten Hindi Text. International Journal Of Computer Applications (0975 – 8887) Volume 1 – No. 4 2010

[10] Naresh Kumar Garg, Lakhwinder Kaur, M. K. Jindal. The Hazards In Segmentation Of Handwritten Hindi Text. International Journal Of Computer Applications (0975 – 8887) Volume 29– No.2, September 2011.

[11] Naresh Kumar Garg, Lakhwinder Kaur, M. K. Jindal. Recognition Of Offline Handwritten Hindi Text Using SVM. International Journal Of Image Processing (IJIP), Volume (7) : Issue (4) : 2013.

[12] R. Jayadevan, S. R. Kolhe, P. M. Patil, Umapada Pal. Database Development And Recognition Of Handwritten Devanagari Legal Amount Words. International Conference On Document Analysis And Recognition 1520-5363/11 IEEE 2011

[13] RAJIV KUMAR, AMRESH KUMAR, PERVEZ AHMED . A Benchmark Dataset For Devnagari Document Recognition Research. Recent Advances In Telecommunications, Signals And Systemsisbn: 978-1-61804-169-2

[14] Sandip N.Kamble, Prof. Megha Kamble. Morphological Approach For Segmentation Of Scanned Handwritten Devnagari Text. International Journal Of Computer Sci Ence & Technology Vol. 2, Issue 4, ISSN : 0976-8491 (Online) | ISSN : 2229-4333(Print) Dec. 2011

[15] Suryaprakash Kompalli • Srirangaraj Setlur Venu Govindaraju. Devanagari OCR Using A Recognition Driven Segmentation Framework And Stochastic Language Models. Springer IJDAR (2009) 12:123–138 DOI 10.1007/S10032-009-0086-82009

[16] Veena Bansal And R. M. K. Sinha. Segmentation Of Touching And Fused Devanagari Characters.

[17] Veena Bansal And R. M. K. Sinha. Segmentation Of Touching Characters In Devanagari.

[18] Vijaya Rahul Pawar, Arun Gaikwad. Multistage Recognition Approach For Offline Handwritten Marathi Script Recognition. International Journal Of Signal Processing, Image Processing And Pattern Recognition Vol.7, No.1 (2014), Pp.365-378 Http://Dx.Doi.Org/10.14257/Ijsip.2014.7.1.34 2014