Handwritten Marathi Compound Character Recognition by Wavelet Technology and Multistage Classifier

Farheen Adil Talib[#1], Prof Dipti Patil[*2]

[#]*Information Technology, Mumbai University*
Konark Plaza, A-wing ,Rm 33Bandar Road, Old Panvel ,410206, India
[1]farheenshaikh85@gmail.com

[*]*Information Technology*
[2]*dypatil75@gmail.com*

***Abstract*— The handwritten character recognition plays an important role in the modern world. Hitherto, any handwritten or printed document, if it is to be replicated digitally, needs to be photocopied or scanned. Such a replicated document cannot be altered in terms of the words, font style and size that the document contains. This paper presents an approach for recognition of unconstrained Compound Marathi character. Multistage feature extraction and classification is used for recogonition of unconstrained characters. Structural feature is used for initial stage and wavelet technology is used for final extraction. The final stage of feature extraction employs generation of kernel. Approximate coefficients are generated by single level wavelet decomposition. This generated coefficients are stored as kernel to perform template matching.**

.

***Keywords*—Compound character, multistage wavelet, convolution, cropped**

I. Introduction

Marathi is an Indo-Aryan language spoken by about 71 million people mainly in the Indian state of Maharashtra and neighbouring states. Marathi is thought to be a descendent of Maharashtri, one of the Prakrit languages which developed from Sanskrit. This work is based on invariant moments for recognition of isolated Marathi Handwritten Characters and their divisions. Handwritten Marathi Characters are more complex for recognition than corresponding English characters due to many possible variations in order, number, direction and shape of the constituent strokes. Indian handwritten characters are more complex due to their structure, shape and presence of modifiers and compound characters. For the computer, it is very difficult to recognize the handwritten character for its disoriented patterns, presence of unwanted object or pattern distortion effect. Handwritten character recognition aims at converting handwritten characters in images into text that can be stored, edited or can be converted into speech. So the challenges are to extract the efficient features from the different characters images of several characters so the different characters can be easily identified by the system. The recognition of handwritten character is one of the major areas of research and is gaining much attention now. A feature extractor and classifier may recognize a character which may not be recognize by other feature extractor and classifier combination. Hence a multistage system is needed that can recognize the characters over a wide range of varing conditions. Here an attempt is made for unconstrained handwritten Marathi compound character recognition without separation of the characters in the compound character.

A. Marathi Script and Compound Characters

Marathi script is derived from Devanagari. Marathi script consists of 16 vowels and 36 consonants making 52 alphabets. Marathi is written from left to right. Marathi also has a complex system of compound characters in which two or more consonants are combined forming a new special symbol. The compound characters in Marathi script exhibit following features: The compound characters in Marathi are formed by joining two or more specific consonants. Fig.1 shows some examples of compound characters in Marathi script that are formed by joining two or more consonants.

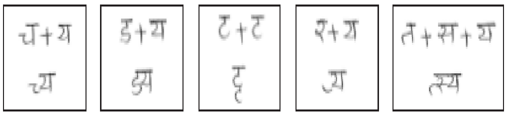


Fig1: Example of Compound character

B. Problem Statement

The objective of this proposal is to provide a novel approach for recognizing unconstrained Handwritten Marathi Compound characters. Here, recognition is carried out using multistage feature extraction and classification. At first the character are scanned. The scanned document is saved in the form of image and this image is processed .Before feature extraction some pre processing is carried out to remove the unnecessary part. After pre processing, a two level structural classification is done using the structural features. Features are classified as global features and local features. Structural feature is used at initial stage and kernels are generated at the final stage of feature extraction and wavelet transform is used at final stage. Here wavelet decomposition is used to generate approximate coefficients. A recognition is carried out using template matching. The output is in the digitized form of recognized text document.

One more important objective of this project is to give the output in the digitized form along with edit facility i.e. the output is editable. Simply we can copy paste the output to the word and make required changes easily.

C. Scope of the Project

In future complete OCR can be designed which will take care of the text with all modifiers and joint characters also and the same can be needed for variety of applications. Some of them are explained below.

One unnoticed application of this work is searching a particular string from a huge document. Consider a case that there is a book whose digital copy is not available on the internet and we want to search some topic from such a book then it will definitely take much more time and the work will be very tedious as book contains around 1000 pages. To overcome this problem OCR can be used. Once we convert whole book in to the digital format, we will be able to search a string within fraction of second as we do now by using Ctrl+F, that is shortcut key to find a string. Due to some awkward words of Marathi on particular page (Words that are difficult to understand for common user), we need focus on OCR is to edit particular page.

II. PROPOSED SYSTEM

The proposed system is design to recognize Marathi Handwritten Compound character. At first the handwritten characters are scanned and stored in png file format. After that the scanned text is stored in the ASCII or UNICODE format. Handwritten or printed character recognition is an important filed of Optical Character Recognition (OCR). The objective of this project is to propose a system for unconstrained handwritten Marathi compound character recognition without separation of the characters in the compound character. The picture of Marathi compound character can be converted to word format for further use. In the proposed system preprocessing is also done to remove the unnecessary part. The scanned document is saved in the form of image and this image is used for recognition. The output of this work is the handwritten compound character that is in the form of recognized text document and is available to user in editable format. The output of this work is the digitized form of recognized text document.

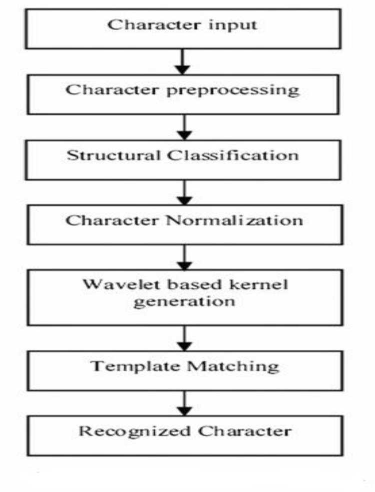


Fig 2: Proposed System Design

1. The proposed system is designed to recognize 40 Marathi compound characters as well as handwritten characters. In Fig. 2 a block diagram depicts the flow of the proposed system.
2. At first, the handwritten compound characters are scanned at 300 dpi in bmp file format. Before feature extraction, some preprocessing is carried. An averaging filter prior to binarization bridges small gaps. The unwanted strokes are further removed by performing morphological operations on the binarized image.
3. After preprocessing, a two level structural classification is done using the structural features. Features are classified as global features and local features. These are the structural features of the character. Global features are used to classify the characters coarsely. Further local features are extracted which further classify these classes, again based upon the structural features which are more character specific than the global features. After the second stage classification, the character is normalized for kernel generation. Single level wavelet decomposition is used to generate the kernels from the resized character images. The recognition is carried out using template matching.

III. IMPLEMENTATION

A. Data Collection

The handwritten compound characters are scanned at 300dpi in bmp file format. The database for handwritten compound characters is created by scanning the characters at 300 dpi using a flatbed scanner. The images are stored in png file format. Here we assume that the consonants in the compound characters are either touching or overlapping. However, sometimes they may not be touching or overlapping due to the writing style or may result into a gap or separate after binarization and preprocessing operations. These split components of a compound character result into separate entities after segmentation. In order to take these separate entities into account, we also consider the split components of the compound character for recognition. No standard database for compound characters is available.

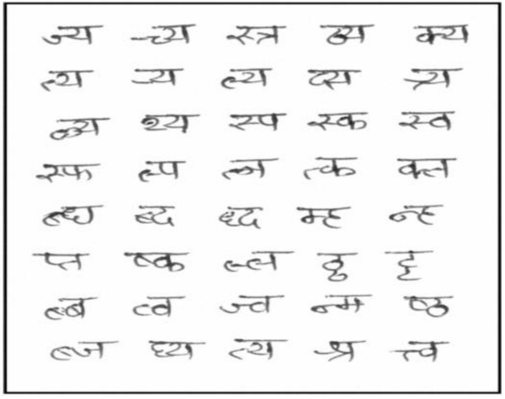


Fig 3: Characters used in proposed system

B. *Structural Classification*

After preprocessing, a two level structural classification is done using the structural features. Features are classified as global features and local features. These are the structural features of the character. Global features are used to classify the characters coarsely. The large number of compound character set with a wide range of variations in the writing style demands a pre classification of the characters before the final recognition. The pre classification is done using a two stage classification based upon the structural features. The first stage employs classification using global features like presence of vertical line in the character, its position in the character and the presence of holes. These features can be termed as global features. The detection of global features is followed by the detection of the local features. The detection of these features is explained in the next section.

The features are classified as global and local features.

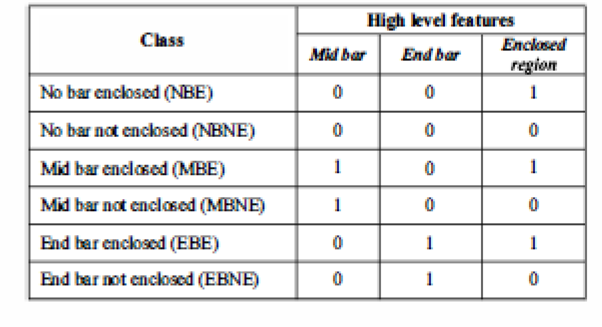
1) Detection of global features

Global features used for classifying the characters at the first stage include presence of vertical bar in the character, position of the vertical bar and presence of the enclosed region. To detect whether these features are present in the character, the following algorithm is implemented. A character image is cropped in order to remove the portions without the character regions. Vertical projection profile of the cropped image f(x, y) is further calculated in order to find the column with maximum number of pixels Y_{max}' . An average height of the vertical bar is considered to be 85percent of the total height of the image. This value is set as a threshold Tv to find the presence of a vertical bar in a character. Thus if, Y_{max}>=Tv vertical bar is said to be present, else, there is no vertical bar in the character. If the presence of vertical bar is detected, further its location is found so as to further classify the character as per its location within the character. Again an average threshold T_M is set to be 30 percent, for the position of the vertical bar in the character. If, T>= T_M, the vertical bar is towards the center else towards the end, where,

$$T = \frac{y - y_{max}}{y} \times 100$$

Further, 8-adjacency is used to find the presence of connected components or the enclosed regions within the character. Two foreground pixels p and q are said to be connected if there exists an 8-connected path between them, consisting entirely of foreground pixels. Table I indicates the classification of the characters based upon these high-level features.

TABLE I
FIRST STAGE STRUCTURAL CLASSIFICATION



Class	High level features		
	Mid bar	*End bar*	*Enclosed region*
No bar enclosed (NBE)	0	0	1
No bar not enclosed (NBNE)	0	0	0
Mid bar enclosed (MBE)	1	0	1
Mid bar not enclosed (MBNE)	1	0	0
End bar enclosed (EBE)	0	1	1
End bar not enclosed (EBNE)	0	1	0

2) Detection of local features

The cropped binary image f (x, y) is thinned to yield a single pixel wide character. This character is then passed to hit-or-miss transformation to find the endpoints of the character. The image is then partitioned into four quadrants as shown below

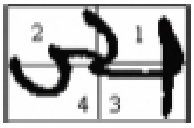


Fig 4: Character Partitionning

Here quadrants 3 and 4 only are of interest. Based upon the presence of the endpoints in quadrants 3 and 4, the six classes obtained in the previous step are further divided into four classes as explained in Table II.

Table II
Second Stage Structural classification

Class	Mid level features	
	End point in quadrant 4	End point in quadrant 3
00	Absent	Absent
01	Absent	Present
10	Present	Absent
11	Present	Present

At this stage, after the classification of characters based on two stage classification using structural features, the characters are classified into 24 classes. For example, a character classified to class EBNE 01 has a vertical bar at the end, without an enclosed region and an end point in quadrant. Further local features are extracted which further classify these classes, again based upon the structural features which are more character specific than the global features. After the second stage classification, the character is normalized for kernel generation. Single level wavelet decomposition is used to generate the kernels from the resized character images. The recognition is carried out using template matching.

C. Generation of Wavelet Kernel

After two stage structural classification, the character is resized to a fixed size of MxN to generate the kernels. Wavelet transform is used for kernel generation. The wavelet transform exhibits the features like separability, scalability, translatability, orthogonality and multiresolution capability. Single level wavelet decomposition is used for kernel generation. The discrete wavelet transform can be implemented using digital filters and down samplers. The high pass or detail component characterizes the image's high-frequency information with vertical orientation; the low-pass, approximation component contains its low-frequency, vertical information. The approximation coefficients obtained for every character after single level decomposition is stored as the kernels in the database. . A modified wavelet kernel generation method is also implemented in order to improve the recognition results. Here the kernels convolved before storing in the database. The approximation coefficients obtained using single level wavelet decomposition are convolved with themselves for generation of modified wavelet kernels .

D. Recognition Technique

Here template matching is used for recognition. The basic template matching technique performs cross correlation of the 2D function f(x, y) with the template g(x, y). The result contains peaks at the location of the matches between the template and the underlying object.

$$r = \frac{\sum_x \sum_y (f-\bar{f})(g-\bar{g})}{\sqrt{\left(\sum_x \sum_y (f-\bar{f})^2\right)\left(\sum_x \sum_y (g-\bar{g})^2\right)}}$$

where, $\bar{f}$ and $\bar{g}$ are the mean of the 2D function and the template respectively and r is the peak of the cross correlation function that indicates the amount of matching. As r moves from 0 to I, the match between the function and the template goes on increasing. The kernels obtained by both the methods are matched with the test character kernel obtained in the same way. The matching is done with the characters in the class to which the candidate character is assigned after structural classification.

IV. RESULTS

The handwritten Marathi compound character dataset is collected. No standard database is available for handwritten Devanagari compound characters.

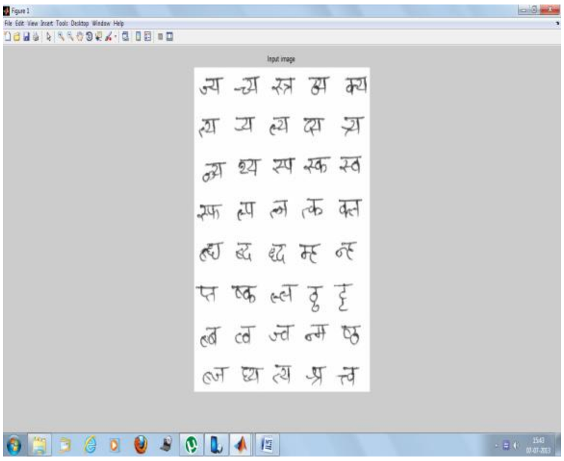
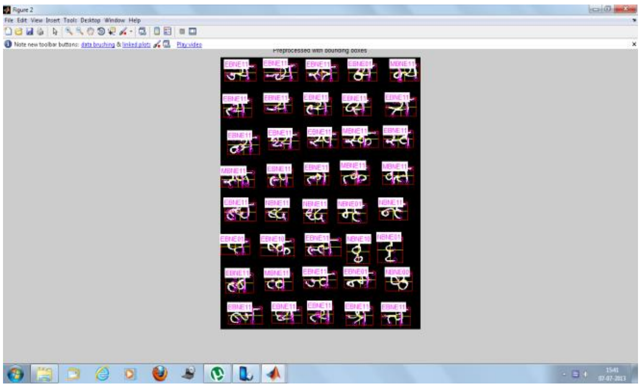
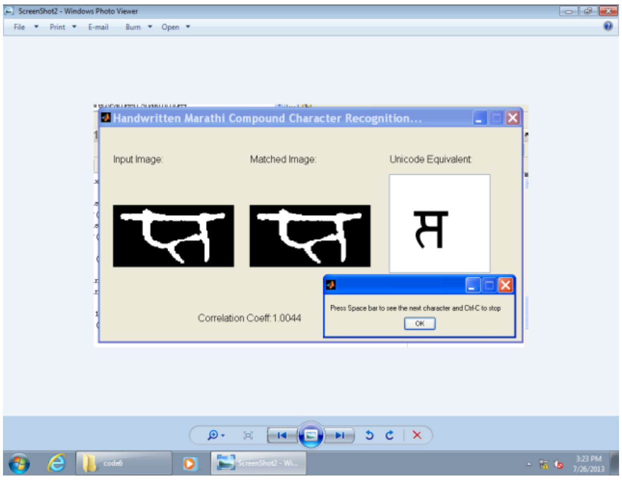


Fig 5: Handwritten compound characters

At first the global features are detected and the characters is classified into six classes. At the second stage the six classes obtained are further classified into four classes.




ScreenShot2 - Windows Photo Viewer
File Print E-mail Burn Open

Handwritten Marathi Compound Character Recognition...

Input Image:

Matched Image:

Unicode Equivalent:

म

Correlation Coeff: 1.0044

Press Space bar to see the next character and Ctrl-C to stop

OK

Fig 7:Recognized image at the output

V. CONCLUSION

After implementation we are able to recognize a document in handwritten using wavelet technology. An attempt is made to design the OCR for handwritten Devanagari script. The handwritten Marathi compound character dataset is collected from more than 5 writers. Out of these two third of the characters are used for kernel generation for the database and the rest are used for testing. No standard database is available for handwritten Devanagari compound characters so far. The handwritten characters may take different shapes as per the writing style of the writer. This may result in classification of the same character to different classes. We propose a novel approach based on multi stage feature extraction and classification scheme for unconstrained handwritten Marathi compound character recognition. The preprocessed character passes through a two stage structural classification. The structural classifier extracts global and local features and classifies the character to one of the 24 classes at the output of the second stage of structural classification. Further, characters are resized for kernel generation. A wavelet based and a modified wavelet based technique is used for kernel generation. The kernels are generated at various resize factors and the results are analyzed after template matching. After implementation we are able to recognize a handwritten Marathi compound character using Wavelet technology and to make text available in editable format for future use

VI .APPLICATIONS

1. Character recognition can be used in bill processing system

2. Assigning pin-codes in postal addresses.

3. Large-scale data processing such as postal address reading check, sorting, office automation for text entry, automatic inspection and identification Character recognition can be used in bill processing system

4. Office Automation.

5. Automatic scoring of tests containing multiple choice questions.

6. It provides a great compression and efficiency by converting the document image from any image file format into more useful formats like HTML or various word processor formats.

7. In job application form sorting

8. In India huge volumes of historical documents and books (handwritten or printed in Devanagari script) remain to be digitized for better access, sharing, indexing, etc. This will definitely be helpful for other research communities in India in the areas of social sciences, economics, and linguistics.

9. Exceedingly user-friendly with features that can edit, move, resize or duplicate the scanned document.

REFERENCES

[1] Sushama Shelke and Shaila Apte “A Novel Multistage Classification And Wavelet Based Kernel Generation For Handwritten Comound character Recognition,978-1-4244-9799-7/11/$26.00@2011 IEEE,pp193-197.7/11/$26.00@2011 IEEE ,pp 193-197.

[2] Utpal Garain and B. B. Chaudhuri, "Segmentation of touching characters in Printed Devanagari and Bangia Scripts Using Fuzzy Multifactorial Analysis", IEEE Trans. Systems, Man and Cybernetics-Part C: Applications and Reviews, vol. 32, no.4, pp. 449- 459, 2002.

[3] V.Bansal and R M. K. Sinha, "On How to Describe Shapes of Devanagari Characters and Use them for Recognition", Proceesings of the Fifth International Conference on Document Analysis and Recognition, pp. 410-413, 1999.

[4] V. Bansal, "Integrating Knowledge Sources in Devanagari Text Recognition", Ph.D. Thesis, lIT Kharagpur, 1999.

[5] Surojit Saha and Dr Ranjan Parekh,” Recogition Of Handwritten Character Based On WREDF And Neural Network Classifiers” ISSN: 0975-5462

[6] Jayashree Prasad and Dr U.V Kulkarni, ”Trends in Handwriting Recognition” 978-0-7695-4246-1/10$26.00@2010IEEE,pp491-495.