REVIEW ON SEGMENTATION AND RECOGNITION METHODOLOGIES FOR OCR SYSTEM OF DEVANAGARI SCRIPT

***Harsh Shah, **Vina Lomte, *Prathamesh Nale, *Shweta Panchpor, *Tarun Agrawal**

**Student, **Asst Prof*

Dept. of Computer Engineering, RMD Sinhgad School of Engineering, India

ABSTRACT

Optical character recognition (OCR) of Devanagari script characters has been a popular research topic for many years and still a lot of work is yet to be done. The strive to achieve utmost accuracy in recognizing the characters is ever-increasing and a task yet to be accomplished. The complexity of this particular script having very peculiar features has attracted many researchers to contribute in this field and showcase their work. Building an OCR for any script involves various phases, the two of vital importance being segmentation and recognition. In this paper, we will review each of the two in detail, by observing the existing methodologies used and their probable scope of improvement.

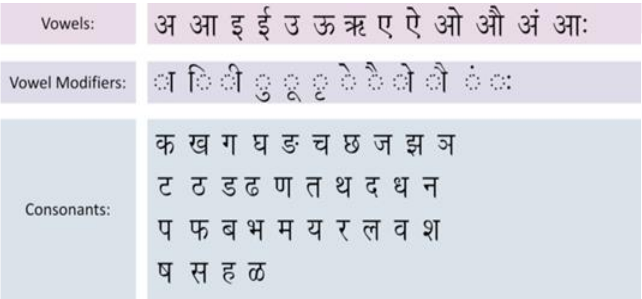
Keywords – *Devanagari Script, Feature Extraction, Optical Character Recognition, Segmentation.*

INTRODUCTION

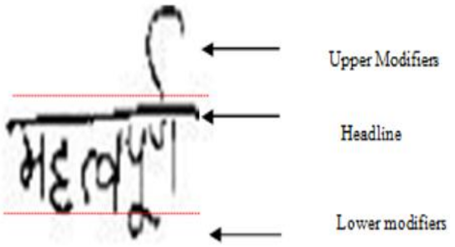
Optical character recognition is getting more and more attention since last decade due to its wide range of application. Although Devanagari being one of the oldest known scripts, a very less research work is done in this field which was successful. The documents related to our history, for example, manuscripts are mostly written in scripts that are difficult to interpret and translate. Conversion of these handwritten characters into machine editable form is important so that it can be easily preserved and accessed to study about the past. A lot of independent work is currently going on in Optical Character Recognition that is processing of printed or computer-generated document and handwritten or manually created document processing for character recognition of various handwritten scripts from the past. In the scope of this paper, we will have an overview of the peculiarities of the Devanagari script and further discuss on the previously implemented methodologies for segmentation and recognition and analyze the research gap. Moreover, a literature survey on segmentation and a methodology survey on recognition techniques will give a detailed comparison on the existing work done in each field respectively.

OVERVIEW OF DEVANAGARI SCRIPT

Devanagari is a script originated from the Sanskrit language, and is used in India since 12th century. Devanagari script forms the base for many languages like Hindi, Marathi, Nepali and Sanskrit. The script is always written from left to right and has a header line which combines each character to form meaningful words. It consists of consonants, vowels, vowel extensions or modifiers and some special characters. The word in a Devanagari script is completed by a header line known as Shirorekha to give a complete meaning to the connected characters. There are 11 vowels along with the 33 consonants the comprise the Devanagari characters. Vowels can be written as independent letters or by using them with consonant they belong to by attaching them above, below, before or after. Vowels when written in this way are known as modifiers and the characters then formed are known as conjuncts. Compound characters can be formed by combining two or more consonants.



Vowels:	अ आ इ ई उ ऊ ऋ ए ऐ ओ औ अं आः
Vowel Modifiers:	◌ा ◌ि ◌ी ◌ु ◌ू ◌ृ ◌े ◌ै ◌ो ◌ौ ◌ं ◌ः
Consonants:	क ख ग घ ङ च छ ज झ ञ ट ठ ड ढ ण त थ द ध न प फ ब भ म य र ल व श ष स ह ळ



OCR FOR DEVANAGARI SCRIPT

The optical character recognition is an upcoming research topic especially when it comes to recognizing scripts that are difficult to read and interpret. Numerous amounts of work have already been done in this field and a lot is yet to be explored. The OCR system for non-Indian languages can be easily developed as the character complexity is less when compared to any character from the Indian script. The development of an OCR system for Devanagari script is indeed a difficult task to accomplish. Various phases are involved when developing an OCR. The two phases of utmost importance and the ones we will further discuss in detail in the scope of this paper are segmentation and recognition.

Character segmentation is a necessary pre-processing step of OCR. The OCR character segmentation is a program that translates a scanned image of a document into a text document that can be edited. It is an operation to decompose the image into sub-images having individual symbols. The accuracy of OCR is directly proportional to accuracy of segmentation. The foremost purpose of this is to provide discrete characters to optical character recognition. An efficient Indian language OCR completely depends on segmentation for better recognition of fused or conjunct characters.

The character recognition of Devanagari script can be done in various ways. The traditional machine learning classification methods are widely used. Although the algorithms available for recognition of Devanagari script characters are limited and not highly efficient due to the character complexity. This promotes a very promising research topic to find algorithms that may classify and recognize these characters with utmost accuracy.

PHASES OF OCR

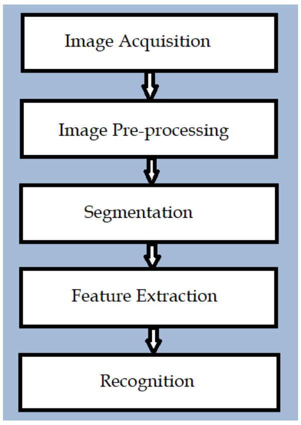


Fig 1

1. Image Acquisition:

The image acquisition or an image input to an OCR involves the process of scanning a document and storing it as an image. Their solution determines the rate of the process and the quality of handwritten text is essential to perform segmentation with efficiency.

2. Image Pre-processing:

The process of preparing the scanned image for segmentation is nothing but pre-processing. It aims to produce data that is easy for the OCR to operate on accurately. Pre-processing may involve various different methods based on the input image to best prepare it for segmentation. It may include binarization, image enlarging, skew correction, grayscale conversion, etc.

3. Segmentation:

Segmentation is one of the two most vital phases of any OCR system. It is the process of decomposing the image into sub-images that will identify individual symbols. Segmentation can be performed using various approaches. Many algorithms are already available to do the same. A lot of already exiting OCR systems have adopted the approach of histogram projections to segment the image into its individual characters.

4. Feature Extraction:

Feature extraction is the process that involves retrieval of the various unique feature a character beholds to distinguish itself from others. Feature extraction is an essential step to help the classifier with the necessary inputs to efficiently recognize each individual segmented character.

5. Recognition:

Recognition is the final phase of an OCR. It involves accurately identifying the individual character. The recognition of characters is usually done using the traditional machine learning classification which is most widely adopted methodology to predict the identified character. Although it is not the only possible way of recognition. Recognition is an untouched and a promising research topic especially for Devanagari text. The state-of-the-art classification can be replaced and we will discuss this in the future scope of this paper.

SEGMENTATION

Any text document contains various words and each word contains many characters. For a recognition system to classify these individual characters they need to be segmented. Segmentation builds the foundation for an efficient recognition system. The recognition of Devanagari characters is an intricate task and hence segmentation plays a vital role. A lot of work has been done in this specific field and a constant research is going on to find new ways of segmenting that can achieve complete accuracy. Some of the work already done by researchers is discussed below.

LITERATURE SURVEY:

In the paper [1] “study of various character segmentation techniques for handwritten off-line cursive words: a review” the authors discuss three segmentation-based approaches for cursive handwriting recognition which they have presented. By the detailed analysis of the literature, it is observed holistic approach is more suitable for applications where the lexicon is statically defined. Explicit segmentation-based approach was computationally complex than implicit segmentation, but gives slightly better results than less complex implicit based segmentation approach.

In the paper [2] “Segmentation of Devanagari Handwritten Characters” proposed algorithms for segmenting Devanagari handwritten script consisting of lines, words and characters are presented. The line segmentation and word segmentation have achieved successful 100% accuracy. But in character segmentation, the accuracy achieved is about 90% this is because the algorithm is not able to properly segment some of the connected words. Besides connected characters the algorithm is efficient in segmenting characters, modifiers and purnviram (full stop in Devanagari).

In the paper [3] “Touching character segmentation of Devanagari script” is it observed that handwritten documents, segmentation of touching characters is very difficult task, the proposed algorithm works well to segment touching characters. The segmentation problem occurs when there is overlapping of ascenders and descenders on two consecutive lines. The same work can be utilized to work with other types of documents such as Hindi, Marathi and other Devanagari languages. The algorithm can be modified to deal with overlapping ascenders and descenders of two consecutive lines as well as two characters overlapping on each other in the same line. This method is unsuccessful for certain characters because of rules we had repressed will not always match to the variation in handwriting i.e., height of two vertical bar characters. As the future work is concern these separated characters as input to the character recognition system.

In the paper [4] “Devanagari Script Conjunct Characters Segmentation Based on Character Structural Properties by Horizontal Projection” it is observed that the algorithms for separating conjunct characters in Devanagari script have shown good result for many conjunct characters. The separation for consonants is basically dependent on the structural properties of the script such as the consonant height difference between the half and full consonants. The consonants or characters are successfully separated in left and right parts. During testing it was observed that sometimes a composite character was substituted by another Devanagari character. The segmentation of conjunct character is about 92 % for the overall algorithm performance.

In the paper [5] “Devanagari document Segmentation using Histogram approach” they have presented a primary work for segmentation of lines, words and characters of Devanagari script. Nearly 100% successful segmentation achieved in line and word segmentation but character level segmentation needs more effort as it is complicated for Devanagari script. This is challenging work due to following reasons. Compound letters are connected at various places. It was also found that in segmentation the difficulty lies in finding the exact connecting points. Upper and lower modifier segmentation needs different approaches. Separating *anusvara* (**.**) and *full stop* (**.**) from noise is critical as both resemble the same. Techniques from natural language processing can be applied here to resolve the problems. Handwritten unconnected compound letter and unintentionally connected simple letter segmentation is also critical.

RECOGNITION

The recognition of handwritten characters has been a research interest for many dues to its wide range of applications. Devanagari being the third most widely used script in the world and having a historical significance it is necessary to recognize the scripts and translate documents from the past. Compared to different scripts unconstrained Devanagari writing is much more complex than the traditional English cursive due to the possible variations in the order, number, direction and shape of the strokes in the script. Due to the presence of multiple loops, conjuncts, upper and lower modifiers and the number of disconnected and multi stroke characters the Devanagari Character recognition is a difficult task to accomplish. The problem of text recognition has been attempted by various approaches. [R1] Template matching is one such approach. This method works effectively for standard fonts but performs poorly with hand written characters. Feature extraction is another method to recognize the characters. It uses statistical distribution of points where each point in analyzed and its orthogonal properties are extracted. The distance of feature vectors of input image is calculated and matched to a maintained database of the same.

METHODOLOGY SURVEY:

Sr. No.	Paper Title	Author	Publication	Year	Methodology			Dataset	Accuracy
					Pre-Processing	Feature Extraction	Classification		
1	A Deep Learning Approach for Optical Character Recognition for Handwritten Devanagari Script	B Dessai, A Patil	IEEE	2019	RGB to Grayscale Inversion, Non-Linear median filtering, Normalization, Skeletonization	Max Pooling	CNN	1750	89.34
2	Recognition of Handwritten Devanagari Characters using	S Shitole, S Jadhav	IEEE	2018	Filtering, Binarization, Dilation	Chain code, Canny Edge Detection, Direction Features	SVM, KNN	4700	94.2 SVM, 89.3 KNN

	Linear Discriminant Analysis								
3	Accuracy Enhancement of Devanagari Character recognition by Gray level normalization	M Jangid, S Srivastava	ACM	2016	Binarization, elimination, Gray level and character Normalization	Gradient local auto-correction (GLAC)	SVM	36172	95.94
4	A scale and rotation invariant scheme for multi-oriented character recognition	N Tripathy, T Chakraborti, M Nasipuri, U Pal	IEEE	2016		Centroid encoding and PCA	SVM	7515	99.25
5	On the Performance improvement of Devanagari handwritten character recognition	P Singh, A Verma, N Chaudhari	ACM	2015	Converting image into two-tone image, Normalization	Standard zoning, elastic zoning, gradient features	MLP	49000	85.11
6	A hybrid feature extraction algorithm for Devanagari script	D Khanduja, N Nain, S Panwar	ACM	2015	Thinning	Combination of structural and statistical features	Neural Network Classifier	22556	93.4
7	Zernike Moment feature for handwritten Devanagari compound character recognition	K.V Kale, P.D Deshmukh, S.V Chavan, M.M Kazi, Y.S Rode	IEEE	2013	RGB to Grayscale, Binarization, Filtering, Normalization, Skeletonization	Zernike Moment based feature	SVM, KNN	27000	98.37 SVM, 95.82 KNN
8	Recognition of non-compound handwritten Devanagari characters using MLP and minimum edit distance	S Arora, D Bhattacharjee, M Nasipuri, D Basu	IEEE	2010	Filtering, Binarization, Dilation	Shadow features, Chain code histogram	Combination of MLP and Minimum edit distance	7154	90.74
9	Performance comparison of features on Devanagari handprinted dataset	S Kumar	ACM	2009	Binarization	Chain code, Kirsch directional edges, gradient distance transform	MLP, SVM	25000	90.116 SVM, 87.86 MLP
10	Handwritten character recognition using elastic machine and PCA	V Mane, L Ragha	ACM	2009	Gray scaling, Morphological Operations, Thinning	Eigen-deformation based features	Elastic-matching based classifier	3600	94.91

Desai et. al [6] used CNN classifier to 89.34% accuracy on a dataset of size 1750 characters. Shitole et. al [7] used Chain code, canny edge detection with gradient features and directional features such as number of horizontal, vertical and diagonal lines for feature extraction. They got 91.2%,90.9%,94.2% accuracy for the three feature extraction techniques for SVM classifier and 81.4%,91.4%,89.3% accuracy for KNN. Tripathy et. al [8] used Centroid encoding and Principal Component analysis (PCA) with SVM classifier to get 99.25% accuracy. Jangid et. al [9] got 95.94% accuracy using Gradient local auto-correction (GLAC) method and SVM classifier. Khanduja et. al [10] got 91.4% accuracy using bot structural and statistical features using MLP with 70 and 40 neurons as classifier. Singh et. al [11] used standard and elastic zoning to accumulate features and then used gradient features like 'Sobel' edge detection algorithm to get 85.11% accuracy for characters using MLP classifier. Zernike Moment based feature extraction technique was used by Kale et. al [12] using SVM and KNN classifiers obtaining 98.37% and 95.82% accuracy respectively. Arora et al. [13] used chain code histogram and shadow features. Combined approach based on Minimum Edit Distance and MLP is used for classification purpose and they obtained 90.74% accuracy. Mane et. al [14] performed elastic-matching based classification technique using Eigen-deformation based features. They used a dataset of 3600 Devanagari characters to obtain 94.91% accuracy Kumar [15] implemented five feature extraction methods, which are chain code, Kirsch directional edges, gradient, distance transform, and directional distance distribution. They used the dataset of 25,000 Devanagari handwritten characters. They obtained 80.6%, 92.4%, 88.1%, 92.0%, 93.5%, and 94.1% accuracy, respectively with SVM classifier and 82.2%, 88.7%, 83.3%, 89.7%, 89.6%, and 91.9% accuracy, respectively, with MLP classifier.

CONCLUSION

This paper presents a review on the existing methods and techniques that have been implemented in the two important phases of building an optical character recognition system for Devanagari script. The various attributes of this script and its complexity of characters which makes it difficult for OCR to achieve complete accuracy are also discussed. Although a lot of existing work can be found related to recognizing Devanagari characters, many have a research gap and are far from building a completely efficient system. There is still a huge scope for new research work in this area and many promising techniques can be used to replace the stagnant existing work which we believe may achieve complete accuracy. This paper will give a glance of the current research done in this area and will encourage new researchers to contribute and come up with unique state-of-the-art methods to procure utmost accuracy.

REFERENCES

[1] Amandeep Kaur, Seema Baghla, Sunil Kumar, "Study of various character segmentation techniques for hand-written offline cursive words: a review" *International Journal of Advances in Science Engineering and Technology*, ISSN: 2321-9009, Volume- 3, Issue-3, July-2015

[2] Ankita Srivastav, Neha Sahu, "Segmentation of Devanagari Handwritten Characters", *International Journal of Computer Applications (0975 – 8887),* Volume 142 – No.14, May 2016

[3] Subith Babu, Mahesh Jangid, “Touching character segmentation of Devanagari script”, *ACM*, ICCCNT '16, July 06-08, 2016, Dallas, TX, USA © 2016 ACM. ISBN 978-1-4503-4179-0/16/07

[4] Ankur Kumar Aggarwal, Aman Kumar Aggarwal, “Devanagari Script Conjunct Characters Segmentation Based on Character Structural Properties by Horizontal Projection”, *IJCST Vol. 4, Issue 2, April - June 2013*, ISSN: 0976-8491 (Online) | ISSN: 2229-4333 (Print)

[5] Vikas J Dongre, Vijay H Mankar, “Devanagari Document Segmentation Using Histogram Approach”, *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT*), Vol.1, No.3, August 2011

[6] B Dessai, A Patil: A Deep Learning Approach for Optical Character Recognition for Handwritten Devanagari Script, 2nd *International Conference on Intelligent Computing, Instrumentation and Control Technologies* (ICICICT 2019)

[7] S Shitole, S Jadhav: Recognition of Handwritten Devanagari Characters using Linear Discriminant Analysis, 2nd *International Conference on Inventive Systems and Control* (ICISC 2018)

[8] N Tripathy, T Chakraborti, M Nasipuri, U Pal: A scale and rotation invariant scheme for multi-oriented character recognition, 23rd *International Conference on Patter Recognition* (2016)

[9] M Jangid, S Srivastava: Accuracy Enhancement of Devanagari Character recognition by gray level normalization, 7th *International Conference on Computing Communications and Networking* Technologies (ICCCNT 2016)

[10] Khanduja, D., Nain, N., Panwar, S.: A hybrid feature extraction algorithm for Devanagari script. 15(1) (2015)

[11] P Singh, A Verma, N Chaudhari: Accuracy Enhancement of Devanagari Character recognition by gray level normalization, *Applied Computational Intelligence and Soft Computing* (2015)

[12] Kale, K.V., Deshmukh, P.D., Chavan, S.V., Kazi, M.M., Rode, Y.S.: Zernike moment feature extraction for handwritten Devanagari compound character recognition. Int. J. Adv. Res. Artif. Intell.3(1), 68–76 (2013)

[13] Arora, S., Bhattacharjee, D., Nasipuri, M.: Study of different features on handwritten Devanagari character. *Int. Conf. Emerg. Trends Eng. Technol.* (2009)

[14] Mane, V., Ragha, L.: Handwritten character recognition using elastic matching and PCA. IN: *International Conference on Advances in Computing, Communication and Control*—ICAC3 ’09, pp. 410–415 (2009)

[15] Kumar, S.: Performance comparison of features on Devanagari handprinted dataset. *Int. J. Recent Trends Eng*. (2009)