Review On Optical Character Recognition for Off-line Devanagari Handwritten Characters & Challenges

Deepu Kumar¹, Divya Gupta²

¹M.Tech Scholar (Computer Science & Engineering), Ajay Kumar Garg Engineering College, Ghaziabad, India
²Asst. Professor (Computer Science & Engineering), Ajay Kumar Garg Engineering College, Ghaziabad, India

ABSTRACT

Off-line handwritten Devanagari script recognition is getting a brighter side of the research day by day. In India, millions of people use handwritten Devanagari script for documentation in northern and central parts of India. The optical character recognition for off-line Devanagari script has been improving day by day. Some innovative steps have been taken into consideration. A bunch of work has been also accounted on handwritten character recognition attempt for several Indian scripts, like Gurmukhi, Gujarati, Oriya, Telugu, Kannada, Tamil, Malayalam, etc. This Off-line handwritten Devanagari script recognition does not have enough reported works. As of late different techniques have been represented by the researchers in the direction of off-line handwritten Devanagari script recognition, many recognition systems for detached handwritten Devanagari characters present in the literature work. The objective of this review paper most desirable feature extraction techniques, as well as classification techniques used for the identification are reviewed in various segments of the paper. An effort is made to address the most crucial consequences reported so far and it is also tried to foreground the better directions of the research to date. This review paper is intended to serve as a guide for the readers, working in the field of off-line handwritten Devanagari character recognition.

Keywords: Handwritten character recognition, Feature extraction & Classification, OCR comparison, Challenges, Applications.

I. INTRODUCTION

Handwritten character recognition is an area of pattern recognition whose purpose is to classify patterns to alphanumeric or other characters. HCR is a mechanism to convert any type of written document file into text format which is editable. Mainly it can be recognized as online and offline character recognition. Off-line Character recognition can be categorized into machine printed and handwritten character. Peoples have a different way of writing, and some have obliqueness in their writing. Such challenges put together the researchers to work out such problems. There is a great need for HCR related research in Devanagari Script, even though there are many challenges as well as the lack of a commercial market [5]. Disseminate of computers in organizations, automatic processing of paper documents is rapidly making importance in India. First attempted work on handwritten Devanagari characters was reported in 1977 but not much off-line handwritten character recognition research work is reported after that. Now the researchers have started working on off-line recognition of handwritten Devanagari characters. A number of research reports are available towards Devanagari numeral recognition but the best of our knowledge there are only a few research reports available on off-line handwritten Devanagari character recognition after 1977 [5].

In this review, we deliver a survey on the different work on OCR for off-line handwritten Devanagari
characters recognition. Most of the reported articles have been reported in the survey done by Pal & Chaudhuri (2004) [1]. We endeavor to deliver a comprehensive survey on handwritten Devanagari characters published from the year (2000) onwards. Our survey binds work like to the handwritten character recognition. A comparison of all the reported feature extraction and classification techniques in the tabular form is referenced. The comparison is done with respect to the features, classifiers, and reported accuracy rate. This analysis showing how the research trend has evolved over the years summarizes the various techniques being applied for classification.

This review paper is formed as follows. Section 1 describes the introduction part. Section 2 describes the basic properties of handwritten Devanagari characters. Section 3 describes the various reported HCR techniques for the script [3]. Section 4 compares the various HCR systems. Section 5 indicating various character –feature extraction approaches. Section 6 includes various classification approaches, future scope in section 7. Conclusion provided in section 8. The references include the most relevant papers recently published as well as some older papers, which can give a comprehensive outline of the developments in the field of the research.

II. PROPERTIES OF DEVANAGARI SCRIPT

Devanagari Script is one of the most famous script within the northern or central division of India and the maximum democratic Indian language [2]. The characters of modern Hindi encompass 13 vowels and 34 consonants. The grouping of vowels and consonants can be named as simple characters. The writing fashion in Devanagari script is from left to right. The common sense of higher/lower case is absent in Devanagari script. In this, a vowel following a consonant takes a changed shape. Depending on the vowel, its modified shape is placed on the left, right or back side of the consonant. These modified shaped referred to as modified characters. A consonant or vowel following a consonant on occasion takes a compound form, which we considered 47 primary characters and the character set considered in this paper appears in figures. The complexity of a handwritten character recognition system will increase particularly because of numerous writing forms of one-of-a-kind individuals [2]. Most of the mistakes in such system arise because of the confusion a number of the in addition formed characters. There are many comparable fashioned characters [4].

The Devanagari script has 13 vowels which are as shown in Fig.1, 34 consonants which are shown in Fig.2, respectively.

![Figure 1. Vowels in Devanagari](image1)

![Figure 2. Consonants in Devanagari](image2)
A different classifier is intended to perceive matras having a place with the upper and lower locale. The characters of center area are perceived in three phases. For center region character classification, every individual character is passed to a three phase classifier to know the correct class of character. For upper area and lower area characters, just single stage classifier has been trained.

Deepki Khanduja et al. [10] in (2016): They have represented the efficiency of any technique is directly proportional on the accuracy of the generated feature set that could uniquely represent a character and hence correctly recognize it. This proposed work included a hybrid mechanism that combined the structural features of character images and a mathematical model of curve fitting to simulate the better features of the character image. Then the combination of structural features of the character such as a number of endpoints loops and intersection points is calculated. This article proposes a model of an efficient feature set extraction technique using statistical and structural features of the text image for script-independent character recognition (it has been tested in English and Urdu also).

Binny Thakral et al. [9] in (2014): Represented a new technique for segmentation of overlapping characters and conjuncts in Devanagari script. The algorithm focuses on the cluster detection technique and works well for various input characters sets. Handwritten Hindi text recognition is a tedious task due to variations in writings, different font size and structural properties of scripts and presence of lower and upper modifiers. They tried to overcome all these challenges and with the help of single algorithm which gives good results with a different type of inputs.

Mahesh Jangid et al. [12] in (2016): They have proposed an HCR for the recognition of handwritten similar shape characters are still a challenging problem in HOCR system. In this article reported an algorithm to estimate the same type of character pairs in Devanagari script and 7 pairs are identified by investigating the confusion matrix. Same shape characters have a minor difference in shape, due to that at the time of classification stage the classifier get confused with another similar shape characters. The solution to that problem by estimated that minor difference known as a critical region in the similar shape characters. The critical region, to extract the more features before classification phase. The critical region is estimated by Fisher discrimination function and a masking technique is used to find the features.

Sushma Shelke et al. [19] in (2015): In this paper, they presented a new approach for the classification of unconstrained handwritten Devanagari characters and performed on the multi-stage classification system. The classification stages categorize the characters into smaller groups. Classification is performed using two stages, the first stage is based on fuzzy inference system and the second stage is based on structural parameters. The fuzzy system improves the classification over crisp classification. The classified characters are communicated to the feature extraction stage. The final stage implements feed forward neural network for character recognition.

Dinesh V. Rojatkar et al. [20] in (2013): In this paper, researchers reported a design and robustness of an optimal classifier applied to off-line handwritten Devanagari characters using single hidden layer feed forward neural network with respect to five-fold cross-validation. Meticulous experimentation of around seventy-five MLPs shows the overall classification accuracy near to 97% for all classes. The best network is found at 5 fold with 80 neurons at trial 3 Networks analyzed on account of confusion matrix, reveals the greater details for individual classes. As a classifier, an MLP NN with log-sigmoid activation functions for hidden layer and output layer gives an impression to perform reasonably. When evaluated for all five folds, Best results are
found on fold 5, the robustness of designed classifier with proposed LRTB features is verified.

Dayashankar Singh et al. [11] in (2015): In this article, they have reported a work for the recognition of handwritten Devanagari characters. The system has been designed with the classifier that is radial basis function neural network (RBF) has been implemented on eight directional values of gradient features for handwritten Devanagari character classification. RBF network with one input and one output layer has been performed to consider the recognition accuracy. Directional group values have been implemented with RBF network which yields good recognition accuracy and reduced training time.

Shailesh Acharya et al.[21] in 2015: in this paper, they delivered a new open images dataset for Devanagari script, Devanagari Handwritten Character Dataset (DHCD). The dataset consisted of 92000 images of 46 unique classes of characters of Devanagari script segmented from handwritten files. They explored the difficulties in recognition of Devanagari characters. Along side, the dataset and proposed a deep learning architecture for recognition of these characters. Deep convolutional neural network (CNN) have shown advanced results to conventional shallow networks in many recognition tasks. They used terms as dataset increment and dropout to prevent the model from overfitting. The proposed system scored most elevated exactness of 98.47% on DHCD dataset.

Table 1. Existing Devanagari OCR systems for handwritten character recognition

<table>
<thead>
<tr>
<th>Authors</th>
<th>Pre-processing</th>
<th>Input pattern</th>
<th>Features</th>
<th>Classifiers</th>
<th>Post-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binny et. al. [9] (2014)</td>
<td>Binarization, Skeletonization</td>
<td>Handwritten character</td>
<td>Cluster detection method, touching and overlapping characters.</td>
<td>Neural Network</td>
<td>Nil</td>
</tr>
<tr>
<td>Dayashankar et al [11]</td>
<td>Binarization</td>
<td>Handwritten character</td>
<td>Gradient features, directional feature group extraction</td>
<td>Radial basis function (RBF) neural network</td>
<td>Nil</td>
</tr>
<tr>
<td>Sushma Shelke et al [19]</td>
<td>Binarization, Normalization</td>
<td>Handwritten character</td>
<td>Fuzzy based</td>
<td>Feed Forward Neural Network (FFNN)</td>
<td>Nil</td>
</tr>
<tr>
<td>Shailesh Acharya et al [21] 2015</td>
<td>Binarization, Normalization</td>
<td>Handwritten character</td>
<td>The raw pixel data generation the beat features</td>
<td>Deep Convolutional Neural Network</td>
<td>Nil</td>
</tr>
</tbody>
</table>
COMPARISON OF PROPOSED WORK

In this segment, we compare the various Hindi OCR systems with respect to the feature set formulated, the dataset and the classifier used, and the accuracy obtained. We have organized the comparison into the table as given below.

This table cites the work, represent the feature set and the classifier used, provide the number of output classes (#Class) considered by the classifier, the size of the training set (#Training) and the size of the test set (#Test), and provide the reported accuracy [5]. The table shows blank entries if the authors have not reported any of these items. For instance, if the classifier used is not trainable then #Training shows a blank entry. Next, we give a summary of the different feature sets and classifiers used by the various OCR methods.

FEATURE EXTRACTION

It’s a process of extracting the usable information from the raw data, which reduces the within-class pattern variability for this purpose, a set of features is extracted for each class to differentiate from other classes [4].

<table>
<thead>
<tr>
<th>Work</th>
<th>Feature set</th>
<th>Classifier</th>
<th>#Class</th>
<th>#Training</th>
<th>#Test</th>
<th>Reported Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satish Kumar et al. [13] (2016)</td>
<td>-feature for middle region character recognition (chain codes, NPW, gradient based, h/v cumulative histograms) -feature for upper and lower region character recognition (NPW)</td>
<td>Multilayer Perceptron (MLP)</td>
<td>--------</td>
<td>3500</td>
<td>--------</td>
<td>93.00</td>
</tr>
<tr>
<td>Deepti Khanduja et al. [10] (2016)</td>
<td>Quadratic curve fitting model</td>
<td>Neural Network</td>
<td>--------</td>
<td>16917</td>
<td>5639</td>
<td>93.04</td>
</tr>
<tr>
<td>Binny et al. [9] (2014)</td>
<td>Cluster detection method, touching and overlapping character.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>94.05</td>
</tr>
<tr>
<td>Dayashankar et al. [11] (2015)</td>
<td>Gradient features, Directional feature group extraction.</td>
<td>Radial basis function (RBF) neural network.</td>
<td>49</td>
<td>1000</td>
<td>1000</td>
<td>95.00</td>
</tr>
<tr>
<td>Sushma Shelke et al. [19] (2015)</td>
<td>Fuzzy based</td>
<td>Feed Forward Neural Network (FFNN)</td>
<td>24</td>
<td>24000</td>
<td>16000</td>
<td>96.95</td>
</tr>
<tr>
<td>Dinesh V. Rojakar et al.</td>
<td>LRTB Feature</td>
<td>Feed Forward Neural Network (FFNN)</td>
<td>32</td>
<td>8224</td>
<td></td>
<td>97.00</td>
</tr>
</tbody>
</table>

In the proposed framework we have actualized pixel cluster identification technique alongside different methods like horizontal profile projection and vertical profile projection system to extract the characters from a word. Pixel cluster detection technique is focused on the way that when two or more characters overlap or touch with one another they structure a group of pixels on the point where they touch and thus this method effectively segments the touching and overlapping characters. The single algorithm has been created to segment the isolated, touching, conjuncts and overlapping characters from a handwritten word composed in Devanagari script.
(Deepti Khanduja et al. [10] in 2015): Structural features are used to take the topological information of the handwritten Devanagari characters, check the minimal effect of shape/font/size variation on the feature set. In this paper reported techniques capture the endpoint, intersection point, and the presence of hash loop in the character image. EndPts are composed of the number of single neighborhood pixels, whereas intersect points are composed of three or more neighborhood pixels. The hash loop feature is true if the character contains one/more closed loop or hole is/are present in the character image.

Gradient Feature Extraction (Dayashankar et al. [11] in 2015): In the proposed work for handwritten Devanagari character recognition with better recognition accuracy, a robust feature extraction technique has become very important. This yields the development of a various feature extraction technique for handwritten character recognition. Gradient feature extraction technique is one of them. In gradient feature extraction, gradients are computed with the help of Sobel operator. There are two templates in Sobel operator named as a horizontal and vertical template.

(Mahesh Jangid et al. [12] in 2016): The gradient direction feature extraction method is more famous due to provide the best features as compare to the other method. After finding the gradient image (contour pixel), there is two distinguished way to estimate the feature vector. The first way is decomposed the gradient image in 8X8 mesh and sum up the magnitude of the gradient of each direction in each mesh region. Another way is used a Gaussian mask or filter in mesh region to find the feature vector.

LRTB features (Dinesh V. Rojatkar et al. [20] in 2013): The handwritten character data set is derived from images obtained by scanning the handwritten consonants in Devanagari script. Thus the dataset shaped, consist of 8224 images of characters (64x64 pixels each) spread uniformly among 32 classes, where each character (class) is repeated 257 times. The raw data (8224 images) stored as a multidimensional matrix acting as a database. This feature extraction method is based on a projection in which LRTB stands for (Left, Right, Top and Bottom) projection are computed outcome in 256 features (64 features from each projection) plus 128 additional features as an addition of individual rows and columns. In order to reduce the computational complexity of MLPNN, the dimensionality of the feature vectors are reduced to 192.

CLASSIFIERS

Classifiers also play an important role in HCR systems. Various different classifiers have been used for Devanagari character recognition.

Feedforward Neural Network (Deepti Khandhuja et al. [10] in 2016): To recognize the character image, the classification stage used the feature vector generated in the feature extraction stage. Classification is accomplished by a two-layer feedforward neural network with back-propagation learning. The topology of the classifier consists of input and output layers with neurons determined by the length of feature vector and number of classes, respectively. The numbers of neurons in the first and second hidden layer are 70 and 40, respectively. Feedforward neural network employs the gradient descendant rule in an attempt to minimize the squared error between the network output values and the target values for these outputs. The trained network is used to test an input character image.
entered by the user and the classifier maps any input pattern to a number of classifications.

Radial Basis Function Neural Network (Dayashankar et al. [11] in 2015): The Radial Basis Function network with one input and one output layer has been used for the training of RBF Network. Radial basis function has been compared with the result achieved in previous related work i.e. Backpropagation Neural Network (BPNN). The comparative result shows that the RBF with directional feature provides slightly less recognition accuracy, reduced training and classification time. In this review paper, a number of samples of two Hindi characters have been taken. Therefore, there are only two outputs at the output layer. The outputs of output layer Y1, Y2 represents whether the character has been correctly recognized or not.

(Dinesh V. Rojatkar et al. [20] in 2013): In the designed MLPNN classifier one hidden layer is used in which processing elements varies from 64 to 128 with the step of 16. In general, it is understood that a number of processing elements are required to solve the complex problem of pattern matching and classification. It is to be reminded that the heart of the classifier is the features used. If features used are reasonably representing the pattern, the variation in performance due to change in other parameters like PEs should be minimum. This subsection demonstrates the variation in performance parameters with respect to change in PEs in the hidden layer. Variation in performance with respect to changes in PEs from 64 to 128 with little deviation of 2.31%, average percentage classification accuracy varies from 94.59% to 96.90% for validation data set.

Multi-Layer Perceptron (Satish Kumar et al. [13] in 2016): In the proposed work, to recognize handwritten Devanagari character a multilayer perceptron has been used. The design used for the neural network is arranged in layers so the model is termed as multilayer perceptron. It composed of a layer of input nodes, one or more layers of hidden nodes, and one layer of output nodes.

Deep learning based system (Shailesh Acharya et al. [21] in 2015): Deep neural networks do no longer require any characteristic to be explicitly defined, instead they work at the raw pixel data generated the high-quality features & used them to classify the inputs into unique classes. For classification, they used the deep convolutional neural network. From this paper, it is noted that the troubles in deep learning based recognition totally popularity because of extra computation time and had much less accuracy. The huge and deep structure of Deep convolutional neural network with a massive bank of trainable parameter makes it vulnerable to overfitting. While training deep deep networks, it is very hard to find optimal hyperparameters of the functions that share the parameters. Those networks being massive require a large amount of training statistics. The to be had dataset in DHCD might not be sufficient to train a community of this length. Few approaches they used to save your the version from overfitting including dataset increment and dropout. These anticipating techniques are not productive to overcome overfitting still their requirements to enhance the framework.

IV. FUTURE SCOPE

In this off-line handwritten Devanagari character recognition great need of research for future researchers. Handwritten Devanagari character recognition having an interesting area of research is word spotting and that would be helpful in large-scale indexing or search in the document images of handwritten archives. Some research is really required to find ideal combinations of classifiers for the purpose of recognition of the handwritten characters. It is still not light up that how a combination strategy can fully utilize the power of sub-classifiers, and to deal with the tradeoff between combination and effectiveness and by hybrid
mechanism and a good level of optimization can generate efficient results which are useful for future OCR. The digitization for better access, sharing, indexing the huge content of historical documents and books written or printed in Devanagari. This will definitely be useful for other research communities in India.

V. CONCLUSION

Evolutionary techniques on different sets of the feature. In this review paper, we have presented various works on off-line Hindi handwritten characters. We have designed the review around works related to handwritten characters and compared the techniques being used in the modern OCR system shown in table 2. Different datasets used in HCR systems. As we realize that each proposed classification framework has it's own particular advantages and disadvantages, yet here we examined those frameworks through perfrormance parameters like exactness, accuracy and affectability and so forth and we characterized significance of each proposed framework through table 1,2. We noticed that in this unconstrained character order there exist heaps of issues to perform on different HCR frameworks. There were potential issues with the current frameworks because of unconstrained characters, improper division of character image, picture procurement issues like ineffectively filtered pictures. Determination of improper element may influence the exactness of the framework. Diverse user writing style and so forth. Incongruent techniques influenced the execution and increase the cost of the framework and similiar character structure appearance influenced the execution of the classification framework. Character recognition segmentation, normalization, feature extraction, classification can be used in an integrated manner for the high accuracy. Handwritten Devanagari character recognition survey can be extended to different classifier like a multiclass model for SVM.

VI. REFERENCES

[9]. Thakral, Binny, and Manoj Kumar. "Devanagari handwritten text segmentation for overlapping and conjunct characters- A proficient technique." Reliability, Infocom Technologies and Optimization
(ICRITO)(Trends and Future Directions), 2014 


