CLASSIFICATION TECHNIQUES FOR HANDWRITTEN CHARACTER RECOGNITION

ABSTRACT - In this paper, we present various classification methods for printend optical character identification (POCR) similarly handwritten optical character identification (HOCR). Observation response of various method illustrate that which scheme produce better recognition trueness in printed optical character identification (POCR) similarly handwritten optical character identification (HOCR). This article illustrate analysis of previous paper, and also distinguish the most important once out of the diversity of superior existing classification and feature extraction techniques and we will standardize the techniques by their feature circumstances and dataset used by different authors. It bring us to the performance of the algorithms produced to the expected efficiency.

Feature withdrawal supporting to examine the shape controlled in the outline. While a quantity of feature taking out and categorization techniques are accessible, other than the picking of an exceptionally superior; technique decides the high degree of recognition correctness. A batch of author present investigation in this field and design novel techniques of extraction and categorization. The core purpose of this proposed article is to re-examine these techniques, so that the group of these techniques can be comprehended.

Keywords: OCR, HCR, Features Extraction, SVM, KNN, PNN, KNN, CNN, ANN etc.

I. INTRODUCTION

Devanagari script - Devanagari is an ancient Indian script. More than 100 languages is written with the use of this script and this script is also used for spoken in India and Nepal including Sanskrit, Hindi, Marathi, Bhojpuri, Nepali, Kokani, Sindhi, Marwari, Maithili, Pali and many more. The word “Devanagari” is a grouping of two words resources God and “Nagari” it depict urban establishment. Thus, “Devanagari” words mean “Script of the God”. It consists with of 49 primary alphabets, 13 vowels, and 36 consonants, and 10 digits also furthermore, the alphabets are modified when a vowel is added to a consonant. This script also helps to write various compound characters. Due to the writing complexity Devanagari script recognition is a challenging task and it has been an active area of research till now[1].

II. OPTICAL CHARACTER RECOGNITION

Optical Character recognition is for the most part of attractive and tremendously attractive region of pattern recognition and artificial intelligence. It is a skill of finding segments and determining lettering from relevant input picture and change it keen on American Standard Code for Information Interchange (ASCII) or other corresponding machine editable and understandable appearance. It participate a part to the improvement of computerization route and getting better the boundary among man and apparatus in a lot of applications. Conversion of handwritten characters is very important in our daily life for construction numeroussignificantpapers related to our olden era, like from manuscripts into machine editable format, so that, that can be merely accessed, processed and store upself-determiningfunctioning on Optical Character Recognition. [2]. OCR is a most important contribution given by progression in computer science and information technology to the mankind. It has finished a lot of deadly work effortlessly and at high rate with enlarged correctness. The a variety of phases consist with OCR technique are:

i. Digitization/scanning/ Acquiring the image
ii. Pre-processing
iii. Segmentation
iv. Feature Extraction
v. Classification &
vi. Post processing

i. DIGITIZATION/SCANNING – Digitization is the process in which conversion of information from electronically into a digital shape. In this system, in a row of data is prearranged into distinct units of information (referred to as bits) that can be disjointedly addressed. similarly Numerical and character images be able to be performed digitally using scanning device that acquire an image from image of text and perform conversion it to an figure files, like as in a bitmap. An optical character recognition (OCR) is application tool that observed
a Numerical and character image for glow and dim areas in order to classify each alphabetic letter or numeric digit, and converts each character into an ASCII code.

**Acquiring the image** - Acquiring the image is a process of producing a digital image from a paper envelope. Either a CCD cameras or a scanner can be used to perform this process.

ii. **PRE-PROCESSING**
Pre-processing is the preliminary segment of HOOCR and it's essential for high-quality identification speed. Stroke normalisation is the main target of pre-processing steps and is remove variations that would otherwise make difficult identification and reduce the recognition accurateness. Such discrepancy or distortions depict the asymmetrical size of a numeral or character image, missing points during pen movement, left or right bend in handwriting and uneven distances of points from neighbouring positions. Pre-processing consist with several stages like normalization of size, resize and centring, interpolating missing points, smoothing, slant correction, erosion and dilation [3].

iii. **SEGMENTATION**
When a task of sub-dividing an Image into constituent parts, or isolating certain aspects of an image is to be referred as segmentation. It also called a process which establish the ingredients of an image, which is compulsory to establish the regions of the paper where data is printed and distinguish them from others. [4]

There are three types of image segmentation.
A. Line segmentation
B. Word segmentation
C. Character segmentation

iv. **FEATURE EXTRACTION**
Feature taking out is individual of the significant segment in identification process and also referred as heart of HOOCR system similar to printed OCR system. It is also determined like extracting the most useful in a row from the raw data(image), which reduced the group outline changeability while attractive the between class pattern variability. A special form of Reduction of information is called Feature extraction method. It reduces the data when input algorithm is very large. [5]

Feature extraction methods are broadly classified as
A. Global Transformation and Series Expansion
B. Statistical Features
C. Structures feature extraction [6]

v. **CLASSIFICATION**
In Optical character recognition Classifiers also play an important role. It is implemented when input image data are presented to OCR system, then features of input image data are extracted and given as an input to the trained classifier. Classification process measure up to the input feature with stored pattern and find out the best matching class for input. Various different classifiers have been used for optical character recognition, like artificial neural network or support vector machine

vi. **POST PROCESSING**
For the purpose of correcting misclassified results by applying linguistic knowledge is refers to Post-processing in OCR. Post-processing is processing of the output from structure recognition. Language information can increase the accuracy obtained by accurate structure recognition. For handwriting input, some shape recognizers yield a single string of characters, while others yield a several number of alternatives for each character images, often with a measure of confidence for each alternative [7]. The whole process described in figure-1

Figure-1 steps of general OCR system

III. **THE IMMENSE IMPACT OF OCR**
OCR (along with paperless document management or computer readable form of data) can make life easier in many ways. Here are a few of the major benefits we stand to increase from the product:

Printed images can be stored efficiently and easily.
Printed material can be easily modifiable once it is kept digital form.

Searching through books or physical form for a specific passage or even a single word simply aren’t practical solutions when working with printed matter.

Improved accessibility and takes less time a book may be accessible to those who are in physical proximity to it.

Safeguarding very important documents becomes much easier when the material is in digital form. Copy of Books, articles and files may all be spoilt beyond repair due to a many of unforeseeable reasons—water leakage, fire, poor handling, etc. Lastly, substantial savings can be achieved by applying OCR and paperless document management. Using advanced skill to convert and scanned print data into digital images lets you avoid spending on filing, archiving, and storage. There is no need for supplies such as paper, ink, or files to store these documents, neither is there any need for human effort to carry out these tasks.

The costs can saved here can work out to thousands over a year.

IV. LITERATURE SURVEY

IV A. ON BANGLA CHARACTER RECOGNITION

Mst. Tasnim Pervin [8] present a paper on feature fusion based optical Bangla character recognition approach using Support Vector Machine (SVM) classification method and the feature extraction method named Zoning and Gabor Filter. This paper presents comparison based on the recognition accuracy by individual features set and by feature fusion which reveals that feature fusion based method performs better (92.99%) than a single feature extraction method (68.15% for Zoning, 89.73% for Gabor filter) during classification process.

Shibaprasad Senet. al [9] this paper, recognition of online handwritten Bangla basic characters has been presented. This paper presents an effort that emphasize the convenience of Hausdorff Distance (HD) and Directed Hausdorff Distance (DHD) based features for the every character sample. They perform Experiment on a set of 10,000 character dataset. Multilayer Perceptron (MLP) has been used for the purpose of classifier that produces the best result with an accuracy of 95.57% when sample character is divided into 16 rectangular zones and DHD-based procedure has been considered for feature extraction.

Shibaprasad Senet. al [10] they present an article on DFA-Based Online Bangla Character Recognition. This paper has been presented investigation on the efficiency of first and second handcrafted feature extraction techniques for the identification of constituent strokes of online handwritten Bangla character samples. These methods determines local and global shape information from a stroke sample.

shared feature vector is passes to Multi-Layer Perceptron (MLP)-based classifier for the purpose of stroke recognition. Finally they have achieved 91.27% recognition accuracy over test set.

MdZahangir Alomet. al. [11] performance of several popular deep convolutional neural networks (DCNNs) for handwritten Bangla character (e.g., digits, alphabets, and special characters) recognition. (experimental results has been presented in this article. They have achieved recognition result of 99.13% for handwritten Bangla digits, 98.31% for handwritten Bangla alphabet and 98.18% for special character recognition using DenseNet.

Shyla Afroleet. et al. [12] they publish a paper on a Hybrid Model for Recognizing Handwritten Bangla Characters using Support Vector Machine. The proposed model is combination of Zernike moments, raw binary pixels and histogram of oriented gradients feature extraction method for recognizing Bangla handwritten characters. Then they have used Support Vector Machine for implementation of classifier. This hybrid system produces 46.98% for Zernike Moments, 66.60% for Raw Binary Pixels and 87.62% for Histogram of Oriented Gradients where overall combined features achieve an accuracy of recognition 94.88%. 6000 dataset has been used for implementation.

IV B. ON MARATHI CHARACTER RECOGNITION

Surendra Pandurang Ramtekeet. et al.[13]This paper present a paper on a Streamlined OCR System for Handwritten Marathi Text Document Classification and Recognition Using SVM-ACS Algorithm and experimental results show that the proposed method has been achieved the accuracy is 99.36 % and the consumption time is 6.55 (ms). This technique is assessed in view of the accuracy, sensitivity, precision, recall and F-score. They have perform a evaluation with the previous Fire Fly Selection (FFS) and Bat Selection (BS) techniques, the proposed method has 99.36% accuracy, 90% sensitivity, 91% precision, 89.51% recall, 99.67% specificity and 89.93% F-score. The proposed methods has actualized using MATLAB and the realtime Marathi character datasets are used for the purpose of experiment.

Parshuram M. Kambleet. et al. [14] present a paper on Geometrical Features Extraction and KNN Based Classification of Handwritten Marathi Characters. They have used 31320 samples of characters from different writers and Noise is removed by using morphological and thresholding operation. using Hough Transformation, Skewed scanned pages and segmented characters are corrected and using bounding box techniques the characters are segmented from scanned pages by Size variation of each handwritten Marathi characters are normalized into different size like 40 X 40 pixel size. For feature
extraction like area, perimeter, eccentricity, orientation and Euler are used. They have used k-nearest neighbor (KNN) algorithm with five fold validation. They have achieved 85.88 % of accuracy by proposed method.


Surendra P. Ramteke et. al. [16] Present a paper on A Novel Weighted SVM Classifier Based on SCA for Handwritten Marathi Character Recognition. This paper explained Various features extraction methods like statistical, global transformation, geometrical and topological features are extracted from the preprocessed image by extraction techniques. they have applied WOAR-SVM classifier perform the best by yielding high accuracy as 95.14%.

IV. C. ON DEVANAGARI CHARACTER RECOGNITION

Richa Sharma et. al. [17] present a paper on Primitive Feature-Based Optical Character Recognition of the Devanagari Script. The presented procedure consist with representation of each character image on the basic of presence and location of primitive features like vertical lines, the frequency, and location of the intersections and the frequency of intersections of character body with Shirorekha of a Devanagari character. The classification of the character is done on the basis of the existence and if available the location of these features in the glyph (test character image). Devanagari handwritten character samples taken from 22 different people from varied age groups for the Ka-Varga—the first five consonants of the Devanagari script. The method worked better for handwritten samples of younger people (aged 20–25 years) than the older ones (aged 40–50 years). They have got 72.72% of accuracy.

Aneeshgupta [18] this article show a concept on neuroplasticity of artificial neural networks: an investigation using english and devanagari character recognition. Author of this paper investigates the neuroplasticity of artificial neural networks for character recognition of English and Devanagari scripts. The neuroplasticity of artificial neural network was trained on images of Devanagari characters, and it recognized Devanagari characters with 97 percent accuracy.

Sarika T. Deokateet. al. [19] Devnagari Script Categorization by Utilizing CNN and KNN, which works well with the categorization of the Devnagari script- Marathi.so it necessitated eradicating the noise and so utilized the Gaussian Approach with Otsu’s approach. At first pre-processing has been performed, then the fragmentation of the content of the manuscript image executed.KNN and CNN categorizers method utilized to categorize the characters, when evaluated on the test and trained dataset with the value of k=1 to 5. The result illustrated for the varying size of the dataset. They have got average 96% of precision for the KNN categorizer.

IV. D. ON GURUMUKHI CHARACTER RECOGNITION

Rupinder Pal Kauret. al.[20] in this article present a concept on Newspaper text recognition of Gurumukhi script using random forest classifier .Author of this paper have presented the recognition performance using random forest classification method for newspaper text printed in Gurumukhi script. They have presented Different kinds of feature extraction techniques to extract the feature of characters after that these values are fed to the random forest classification method with Standard k-fold cross validation and dataset has been used for experimental work. Their proposed method has been produced, maximum recognition accuracy of 96.9% and 96.4%, by the use of 5-fold cross validation and dataset partitioning strategy, accordingly.

Neeraj Kumar et. al. [21] present a thesis on A Novel Handwritten Gurumukhi Character Recognition System Based on Deep Neural Networks. They have proposed work on the feature extraction method using three types of features, namely Local binary pattern (LBP) features in addition to directional features and regional features. In order to correctly recognize the text they have achieved total 117 features to map the Gurumukhi text with Devanagari text a suitable mapping technique has also been implemented. For the purpose of implementation total of 2700 samples have been taken for training and testing purpose. Their proposed system has been produced an accuracy of 99.3% as a performance of system Harmit Kaur et. al. [22].they have present a paper onHandwritten Gurumukhi Character Recognition by using Recurrent Neural Networkthey have used three types of feature extraction namely zoning features, intersection and open end point feature and horizontal peak extent features. For the purpose of classification RNN classifier has been used. In this proposed work they have used, 2450 samples of handwritten Gurumukhi characters and their proposed system has been achieves 87.34% of recognition accuracy.
HarpREET Kauret, el. [23] Handwritten Gurumukhi Character Recognition Using ConvoLution Neural Network they have performed experiment partition strategy for selecting the training and testing sample. 2450 sample images of Gurumukhi characters for the purpose of training and testing has been used. The author have been used Zoning, Diagonal and Horizontal Peak Extent feature extraction techniques for intension to find the feature sets for a given character image. Their proposed system has been achieved a maximum recognition accuracy of 92.08% with 90% training data and 10% testing data using Zoning based features and CNN Classifier.

Rupinder Pal Kauret, et. al. [24] present an article on newspaper text recognition of Gurumukhi script using random forest classifier. they have used a deep learning paradigm using a ConvoLution Neural Network (CNN) for handwritten Gurumukhi and Devanagari character recognition (HGDCR). In experiment of this paper shows the training of a 34-layer CNN for a 35 class self-handwritten handwritten Gurumukhi and 60 class (50 alphabet and 10 digits) handwritten Devanagari character dataset has been performed on a GPU (Graphic Processing Unit) machine. Their experiment putcome with an average recognition accuracy of more than 92% for Handwritten Gurumukhi Character dataset and 97.25% for Handwritten Devanagari Character dataset.

Gupta, Sheifaliet, el. [25] present a paper on Offline Handwritten Gurumukhi Character Recognition System Using ConvoLution Neural Network. This article implemented convolutional neural networks (CNN), for purpose of classifier and then its mapping has been performed to the right observation. They have used a dataset for 10 Gurumukhi characters. Without making use of any post-processing method. Author of this paper proposed methodology that produced a recognition accuracy of 99.34% on Gurumukhi characters images

Rupinder Pal Kauret, et. al. [26] present an article on Recognition of newspaper printed in Gurumukhi scriptFour feature extraction techniques, such as , zoning features, diagonal features, parabola curve fitting based features, and power curve fitting based features has been used for extracting the statistical properties of the characters printed images. For recognition purpose, four classification techniques, namely, k-NN, linear-SVM, decision tree, and random forest has been applied. The author implements random forest and SVM with 5-fold cross validation classifier for recognition accuracy of 96.19% with a combination of zoning features, diagonal features and parabola curve fitting based features. They have achieved a recognition accuracy of 95.21% with a partitioning strategy of data set (70% data as training data and remaining 30% data as testing data) has been achieved.

V. CONCLUSION

In This research work deals with the recognition of Handwritten Gurumukhi, Marathi, Bangali and Devanagari Characters using different classification techniques by implementing combination of Various feature Extraction methods as well as individual method. We have observed that these feature extraction techniques and classification method produced different recognition accuracy. Proposed surveyed work delivers more efficient and highest accurate results on Handwritten Gurumukhi, Marathi, Bangali and Devanagari Character 99.60%, 99.36%, 98.18% and 96.00% accordingly[27].

<table>
<thead>
<tr>
<th>Ref. no.</th>
<th>Author</th>
<th>Year</th>
<th>Methods</th>
<th>Accuracy</th>
<th>Script</th>
</tr>
</thead>
<tbody>
<tr>
<td>[8]</td>
<td>2018</td>
<td>SVM</td>
<td>92.99</td>
<td></td>
<td>Bangla</td>
</tr>
<tr>
<td>[9]</td>
<td>2017</td>
<td>SVM</td>
<td>95.57</td>
<td></td>
<td>Marathi</td>
</tr>
<tr>
<td>[10]</td>
<td>2018</td>
<td>SVM</td>
<td>91.27%</td>
<td></td>
<td>Devanagari</td>
</tr>
<tr>
<td>[11]</td>
<td>2018</td>
<td>DCNNs</td>
<td>98.18%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[12]</td>
<td>2017</td>
<td>SVM</td>
<td>94.88%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[13]</td>
<td>2018</td>
<td>SVM-ACS</td>
<td>99.36%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[14]</td>
<td>2017</td>
<td>KNN</td>
<td>85.86%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[16]</td>
<td>2019</td>
<td>SVM-SCA</td>
<td>85.86%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[17]</td>
<td>2018</td>
<td>Primitive based</td>
<td>72.72%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[18]</td>
<td>2018</td>
<td>ANN</td>
<td>90.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[19]</td>
<td>2019</td>
<td>ANN,KNN</td>
<td>96.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[20]</td>
<td>2019</td>
<td>Random forest</td>
<td>96.90%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[21]</td>
<td>2017</td>
<td>DNN</td>
<td>99.60%</td>
<td></td>
<td>Gurumukhi</td>
</tr>
<tr>
<td>[22]</td>
<td>2016</td>
<td>RNN</td>
<td>87.34%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[23]</td>
<td>2017</td>
<td>CNN</td>
<td>96.08%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[24]</td>
<td>2018</td>
<td>CNN</td>
<td>92.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[25]</td>
<td>2019</td>
<td>CNN</td>
<td>99.34%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[26]</td>
<td>2019</td>
<td>KNN,SVM</td>
<td>96.19%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

VI. ACKNOWLEDGEMENT

This survey paper is based on recently research has been conducted on “classification techniques for handwritten character recognition”. In RNTU, I am grateful to my research guide Dr. Shailja Sharma and HOD Dr. Sanjeev Kumar Gupta to showing confidence in my work and for their patience & support.
I am also thankful to Mr. Rakesh Kr. Roshan, Asst. Prof. RRSDCSE Begusarai Bihar to helping me to write this paper.
I am thankful to my friends, whom never stopped challenging me & helping me to developing my ideas.

VII. REFERENCES


