Comparative Analysis Of Advanced Classification Techniques For Multilingual Ocr Systems

Rohit Verma, Dr. Jahid Ali

Abstract: Classification engineering is reported to be very critical and tedious task in the field of data, image and pattern recognition. Labelling the images into one of the earlier defined categories, is the responsibility of a typical classifier. Preprocessed image does require for decent results. For fantastic and fabulous achievements, image should be free from any kind of noise and should be normalized to the acceptable parameters. There are myriads of classification techniques but the most challenging aspect is to identify the best technique which could intelligently recognize optical characters on the basis of predefined features of characters. This paper elaborates extremely important classification techniques viz. K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Artificial Neural Networks (ANN) and Convolutional Neural Network (CNN). Various classification techniques are compare on the basis of literature so that the researchers could take the advantage and select the best possible technique for their objectives.


1 INTRODUCTION
Classification is defined as a machine learning application of assigning labels to the unknown data on the basis of available set of labeled data. The extremely important task involves the prediction of unknown data based on predefined features. The set of such features which are helpful for recognition of a stranger is referred to as feature vector. For recognition purpose, firstly, a training set of features is developed and then the character to be recognized is processed to produce the same set of the features. These features are compared with the feature vector. The class processing the nearest neighbor is assigned to that character. For attaining high degree of accuracy, choice of features play extremely important role. Most relevant features could improve the accuracy rate to manifolds. Although countless classification techniques are available for achieving the targets, but different classification techniques are meant for different types of classification. In reality classification phase is the critical decision making phase which uses the features extracted in the previous phase to classify the objects.

2 PHASES OF OCR
Following steps could be considered extremely important for the recognition of characters:[1]

2.1 Characters pre-processing
This task includes improve the intensity of characters under observation, Skew removal if any, removal of noise, normalization of the text to the desired level and component analysis of the characters.

2.2 Character detection and extraction
In this step the exact position of the characters is calculated. In case of Devanagari script the characters from the word are segmented from the bottom. Whereas in case of roman characters, they are already segmented. So no significant efforts are employed to achieve the objectives. For the segmentation purpose, initially characters are thinned and then header line is found by knowing the maximum intensity of white pixels. After getting the row number, those columns are set to ‘0’ where only one pixel is found as shown in the Figure.

Fig1. Character Segmentation of Devanagari

2.3 Classifier training
This task is extremely important for the categorization purpose. Most significant features of characters are considered for the training of the classifiers. On the basis of excellent choice of these features a high degree of accuracy could be achieved. The extremely vital features that could be considered for desirable results may be Local Binary Pattern(LBP) Hu’s moment invariants, Zone densities and identification of horizontal and vertical lines in each zone for instance. Other features that could be considered are elaborated in the paper [2]. R Verma et al. presented a set of features and categorized them as per their characteristics.

2.4 Character classification
This is the final step of classification. Important features of the unknown character are generated and compared with pre-generated feature vector of desired domain. On the basis of the match Found, a suitable label is allocated to the character.

3. OVERVIEW OF CLASSIFIERS
Numerous classifiers are available for the recognition of characters for the applications in banking industry, money...
counting, postal department, automatic data entry etc. Majority of classifiers that are employing for attaining the objectives are: k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Local Binary Pattern (LBP). The present work is focused on the comparative study of the most commonly employed classifiers in view of the domain where these classifiers are used.

S. Arora in the paper [3] performed comparative analysis of SVM and ANNs by applying these classification techniques on Devanagari character set. After applying preprocessing techniques, vital features viz. shadow features, view based features, chain code histogram features and longest run features were extracted. The extracted features are input to ANN classifier and in SVM for classification. Desirable accuracy was reported by applying these classification techniques on a count of 7154 data samples of Devanagari script. The most commonly used classifiers are described here:

### 3.1 Artificial Neural Network (ANN)

ANN imitates developing an intelligence that emulate the intelligence of human beings. ANN comprises of cycle of layers from input layer to the output layers in between hidden layers. Each layer has neurons as their fundamental element, which passes information layer to layer. Neurons of various layers are connected by weights to the neurons of the precurasive and subsequent layers. ANN classification technique employs non-parametric approach[1] and the performance of the classifier depends upon the architecture and count of inputs. Anita pal. [4] Reported an accuracy of 94% by applying two-layer perceptron neural network on handwritten English characters. In the paper [5], Manish Kumar et.al reported maximum accuracy of 83% by applying feed forward back propagation neural network on a sample of 297 hand written characters.

![Fig2. Neural Network with six inputs and four outputs](image)

Further it was concluded that recognition approaches depend on nature and count of the sample data under observation. Reetika et.al. in the paper [6] reported an accuracy of 98.77% in the process of recognition of English characters by employing SURF features using Neural Network. The reported SURF Features in addition to neural network has noticeable performance and produces absolutely disjoint set of feature points which enhances the computing speed manifold.

### 3.2 K-Nearest Neighbors (KNN)

KNN is significantly important classification algorithm which has countless applications in Machine Learning (ML). This algorithm is based on the approach of classifying the objects on the basis of closest training match in the feature domain. It is a type of instance based learning, where computation is deferred till classification. Using the approach of k-NN an object is classified on the basis of maximum votes of its k-neighbors and if k = 1, only class is assigned to that object [7]. It is proved to give fabulous results in the domain of supervised learning and finds great applications in pattern recognition field. Though KNN can be used for both classification and regression predictive problems but high degree of accuracy and desirable results are achieved in classification applications.

![Fig. 3 K-NN Classification Diagram](image)

S. Roy in the paper [8] reported an accuracy of 97.67% for classifying the MNIST numeral digits images with KNN and ten features. It is observed that in the domain of pattern recognition KNN is a non-parametric approach used for classification and regression. V. Vaidehi [7] proposed that K-NN classifier produce better and significant results than its other competitors due to reduce error rate and lesser execution time. In addition, this classifier works better with images under different lighting conditions and is independent of image resolutions. In the paper [9], Zanchettin et.al. Proposed a hybrid approach of using K-NN with SVM (Support Vector Machine) for the recognition of hand written cursive English characters and it was observed that the recognition rates improved significantly with this hybrid approach of K-NN and SVM.

### 3.3 Support Vector Machine (SVM)

SVM is one of the most important supervised machine learning algorithms producing excellent results for critical classification and regression challenges. However, it is commonly employed for classification problems to achieve desirable results. Here every data entity is represented as a point in n-dimensional coordinate system and each feature depicts the value of one specific coordinate. After plotting the classification objective is achieved by drawing the hyperplane which could differentiate the two classes without any ambiguity. The distance between the data point and the hyper plane is called margin. SVM best segregates the two classes with hyper-plane but at the same time it could handle multiple continuous and categorical variables also. S.Sharma et al. [10] achieved an accuracy of more than 98% using SVM for bi-class and multi-class character recognition approach. This paper proposed the concept of self-learning. Numerous features were extracted for attaining high degree of accuracy.
K. Gauri et al. in paper [11] reported an accuracy of 97.16% for digits, 95.74% accuracy for capital letters and 92.19% for small letters with SVM classifier. Furthermore, it was reported that SVM employs Structural Risk Minimization principle (SRM) which minimizes an upper bond on the expected risk. In contrast ERM minimizes the error on the training data. Because of this superiority, Support vector machine gave better results. In addition, due to the absence of Local Minima in SVM, it possesses lesser value of execution time in comparison to MLP. Mustafa S. Kadhm et al. in the paper [12], purposed a new architecture based on SVM classifier for recognition of hand written Arabic words. Since the purposed research was based on recognition of word, so character segmentation stage was not required. Further, it was reported that the best accuracy of 96.317% was recorded for the recognition of Arabic Hand written dataset with multiple feature extraction methods and SVM classifier. With the experimental results it was also proved that the polynomial kernel of SVM is convergent and produce excellent results in comparison to other kernels of SVM. In the paper [13], T. W. Umapada Pal, reported an accuracy of 95.13% for the recognition of DevNagari character recognition. To achieve high degree of accuracy two sets of features were used with combined classifiers SVM and MQDF (Modified quadratic function).

### 3.4 Convolutional Neural Network

Convolutional Neural Network (CNN) is extremely important and widely employed technique for pattern recognition. Myriads of reasons have been developed for the famousness of CNN. Traditionally, in character recognition models, feature extraction is is done manually. Whereas in CNNs, the weighted convolutional layer is employed for feature extraction as well as the fully connected layer employed for classification are determined during the training phase. As a reward, memory requirements and computational complexity reduced significantly and performance becomes noticeable for applications where the input has local correlation. A high recognition degree of accuracy is achieved with CNN only. CNN is reported to give a fantastic accuracy of 99.77% testing with MNIST handwritten digits dataset [14]. Furthermore, CNN are shift Invariants because same weight configuration is employed across space. Another important feature of CNN is its noticeably reduced memory requirement. In addition, CNN takes lesser time to train and provides excellent performance. The only limitation of CNN is that it requires tremendous amount of data for achieving desirable and fabulous results.

In paper [15], A. El-Sawy et al. achieved an accuracy of 94.9% in the recognition process of Arabic handwritten characters by Implementing CNN. It was observed that CNN reported best results in contrast to other Deep learning architectures in the domain of image and big data.

![A typical Convolutional Neural Network](image)

In the paper [17], W. Yang et al. proposed an Enhancement of DCNN approach to online Chinese hand written characters by employing domain specific knowledge viz. non-linear normalization, path signature, for instance.

### Table 1: Performance comparison of various classification methods with accuracy

<table>
<thead>
<tr>
<th>S.NO.</th>
<th>Classifier</th>
<th>DataSet</th>
<th>Data Set Size</th>
<th>Accuracy</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SVM and MQDF</td>
<td>Devnagari handwritten characters</td>
<td>36172 characters</td>
<td>95.13%</td>
<td>[13]</td>
</tr>
<tr>
<td>2</td>
<td>SVM</td>
<td>Arabic hand written dataset</td>
<td>1675 words</td>
<td>96.3%</td>
<td>[12]</td>
</tr>
<tr>
<td>3</td>
<td>SVM</td>
<td>Numeric digits</td>
<td>--</td>
<td>97.16%</td>
<td>[11]</td>
</tr>
<tr>
<td>4</td>
<td>SVM</td>
<td>English alphabets</td>
<td>2600</td>
<td>95.23%</td>
<td>[10]</td>
</tr>
<tr>
<td>5</td>
<td>SVM and KNN</td>
<td>Cursive characters</td>
<td>57293</td>
<td>89.65%</td>
<td>[9]</td>
</tr>
<tr>
<td>6</td>
<td>SVM</td>
<td>Assamese character set</td>
<td>--</td>
<td>96%</td>
<td>[18]</td>
</tr>
<tr>
<td>7</td>
<td>Neural Network</td>
<td>Hand written characters</td>
<td>297</td>
<td>83%</td>
<td>[5]</td>
</tr>
<tr>
<td>8</td>
<td>MLP</td>
<td>Isolated hand written Bengali Characters</td>
<td>3000</td>
<td>95.10%</td>
<td>[19]</td>
</tr>
<tr>
<td>9</td>
<td>KNN</td>
<td>Kannada Vowels</td>
<td>--</td>
<td>100%</td>
<td>[20]</td>
</tr>
</tbody>
</table>
4. CONCLUSION

In this paper, widely employed classification techniques for the characters recognition have been reviewed. It is observed that depending upon the dataset used, various classification techniques provide multi-degree of accuracy. In some cases SVM and KNN are proving their presence where as in others CNN is proved to be excellent. In comparison to other classification techniques, CNN gained much popularity because of less memory requirements and reduced training time. But at the same time CNN requires huge amount of data for fabulous performance. Trends have been noticed where CNN achieved an accuracy of 99.77% in case of English characters using MNIST dataset.

REFERENCES


