Classification of the Bangla Script Document using SVM

Manoj Kumar Shukla[1], Ajay Rana[2], Haider Banka[3]
Amity School of Engineering,, Amity University, Noida[1-2]
Department of CSE, ISM (Dhanabad)[3]
E-mail:- mkshukla001@gmail.com [1], ajay_rana@amity.edu [2], hbanka2002@yahoo.com [3]

***Abstract*- In a multi-lingual country like India, identification of the multi-script in an image of a document page is of primary importance for a system processing multi-lingual document. In this paper we present a technique for classify printed Bangla document. Our script classification approach is called SVM based classification. In the classification stage which is the second step of the Optical Character Recognition Engine. . It includes making decisions regarding membership of class of a pattern which is being studied. The aim is to design a decision algorithm that easily computes and minimizes the probability of misclassification. SVM classification system has been tested on printed Bangla Language script and an average accuracy of 94.78% has been achieved.**

Keywords: - Bangla script document, SVM, Feature Extraction, OCR

I. INTRODUCTION

India is a multi-lingual, multi script nation. Consequently creating a fruitful multi-lingual OCR, framework for classification proof of distinctive scripts is a major step. In a multi-lingual nation like India, it is exceptionally vital for outlining an OCR framework. In India 23 official languages namely Hindi, Punjabi Bengali, Maithili, Malayalam, Nepali, Oriya, , Sanskrit, Tamil, Telugu Assamese, Santali, Sindhi Bodo, Manipuri, Marathi Dogri, English, Gujarati, Kannada, Kashmiri, Konkani, and Urdu. There are 13 separate scripts Gurumukhi, Devanagari, Bangla Assamese, Gujarati, Malayalam, Oriya, Roman, Kannada, Kashmiri. Optical Character Recognition (OCR) is the most Important and testing zone of Image Processing & design distinguishment valuable in numerous handy applications like, perusing support for the visually impaired, programmed perusing for sorting of postal mail, bank checks and so on. Our examination we have found some work in Devanagari script division. Manoj et al [1] have used Structural Features Extraction for Devnagari and Bangla language Documents. Manoj et al [2] prescribed a system for classification of the Devnagari Script Document. Manoj et al [3] have proposed a Line-wise Script Segmentation for Indian Language Documents. Manoj et al [4, 7] depict the Degraded Script Identification for Indian Language. Manoj et al [5,] have used the thought of an efficient Segmentation scheme for the recognition of printed Devnagari script. Manoj et al [6, 8] fabricated the clear framework for line base multi script identification for Indian Language.

II. FEATURES OF BANGLA SCRIPT

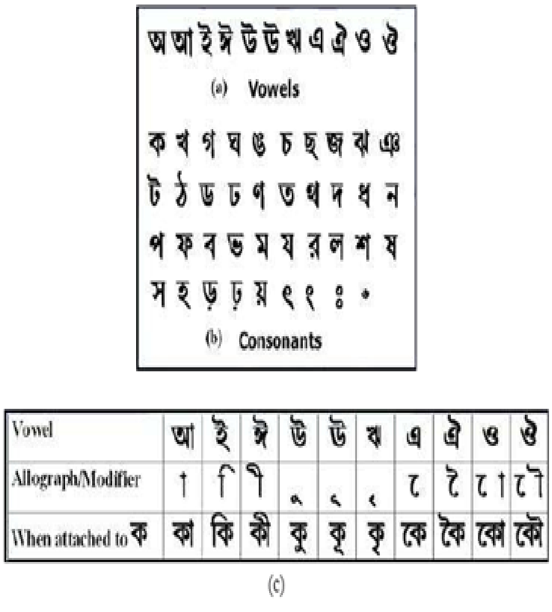
Patterns of Bangla scripts are quite complicated. In comparison to Roman scripts, each word of Bangla sentences consists of various characters clubbed together with a top horizontal line (called 'Maatra' or head-line). It may also contain various compound characters, vowels and consonant signs ('Kaar' and 'Falaa' symbols). Therefore, developing OCR Bangla printed scripts is not easy and requires considerable effort and time.

Prime qualities of any Bangla printed script are as follows:

i. Bangla is written in left to right format.

ii. There is no concept of upper and lower case as in English.

iii. Vowels, among other characters, often take altered forms in a word and named as modifiers or allograph (in Bangla 'Kaar'). Consonant modifiers are also possible (called 'Falaa'). Some of them are shown in following Tables.


অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ
(a) Vowels
ক খ গ ঘ ঙ চ ছ জ ঝ ঞ
ট ঠ ড ঢ ণ ত থ দ ধ ন
প ফ ব ভ ম য র ল শ ষ
স হ ড় ঢ় য় ৎ ং ঃ ँ
(b) Consonants

Vowel	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
Allograph/Modifier	া	ি	ী	ু	ূ	ৃ	ে	ৈ	ো	ৌ
When attached to ক	কা	কি	কী	কু	কূ	কৃ	কে	কৈ	কো	কৌ

(c)

Table-1:- Bangla Character

978-1-4799-8578-4/16/$31.00 ©2016 IEEE

Consonant	Corresponding Consonant Modifier
য	্য
র	্র
র	র্
ঋ	ৃ

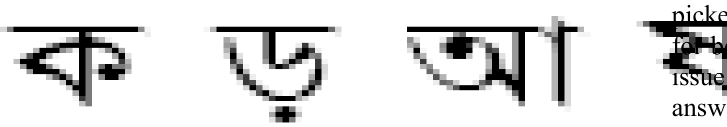
Table-2: Bangla Corresponding modifier forms

- **iv.** Various consonants in one given syllable of a word, may be joined together to make a modified character that preserves the shape of the constituent characters partly (e.g. Na + Da, Ka + Ta, Va + Ra-falaa, Na + Daa + Rafalaa shown in modifier forms).

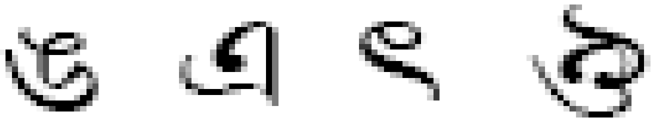
Compound Character	Formation of the Character
ন্ড	ন + ড
ক্ট	ক + ট
ভ্র	ভ + ্র
ন্দ্র	ন + দ + ্র

Table-3: modifier forms

- **v.** Nearly all Bangla symbols and alphabets contains a top horizontal line called 'maatra' with few exceptions (e.g. Ae, Oy, O, Ow, Kha, Ga, Ungo, Nioetc).
- **vi.** Word is composed of the characters that are combined by the virtue of their 'maatra' and other characters and symbols having no 'maatra' (e.g. Khondota, Bishorgo, Ungo, Ae, Oyetc) stay disconnected in the word.



Bangla character with 'maatra' or headline



Bangla characters without 'maatra'

Composing style in Bangla is from left to right. The idea of capital and minor characters is not there in Bangla script. A line of Bangla script could be parceled into three flat zones, specifically, upper, center and easier zone. These three zones are portrayed, with the assistance of one sample word. The center zone usually comprises of the consonants. The upper and more level zones might hold parts of vowel modifiers assistant sings and half characters. In center zone, the majority of the characters hold a flat line on the top, as demonstrated. This line is known as the feature.

III. CLASSIFICATION

Classification is a very import step during identify or recognition of the script. Recently clarified, arrangement is the part of the recognition framework that endeavors to catch the class that a specific character fits in with. After a classifier may do this, it must be demonstrated an expansive number of preparing examples, in a sort of studying stage [10]. It has been noted that the way to high execution is through the capability to select and use the different characteristics of characters. No straightforward plan is prone to accomplish high recognition rate, henceforth more complex frameworks have been created. We have talked over in this area different sorts of arrangement routines that have been investigated. Rundown of different characterization systems incorporates pattern matching, syntactic routines, factual techniques, fake neural systems, bit strategies and half breed classifiers [11-13].

A. SVM

Statistical learning theory is the basis of SVM. It utilizes supervised learning. In supervised learning, a machine is prepared rather than customized, to perform a given errand on various info yield sets. As indicated by this worldview, preparing implies picking a capacity which best portrays the connection between the inputs and the yields. The focal issue in measurable learning hypothesis is the way well the picked capacity sums up, or how well it appraises the yield for beforehand concealed inputs. By and large, any learning issue in measurable learning hypothesis will prompt an answer of the sort.

$$f(x)=\sum_{i=1}^{l} c_i K(x,x_i)$$

978-1-4799-8578-4/16/$31.00 ©2016 IEEE

where *xi* , *i* = 1, …, *l* are the input examples, *K is* a certain symmetric positive definite function known as kernel, and *c*i a set of parameters to be determined from the examples.
The working of SVM can be described as similar to a statistical learning machine that maps points of different categories from *n*-dimensional space into a higher dimensional space where the two categories are more separable. It tries to find an optimal hyper plane in that high dimensional space that best separates the two categories of points.

Basically, the points closest to the hyper plane learn the hyper plane. These points are known as support vectors. More than one support vector may be present on either side of the plane. Figure 1 shows an example of two categories of points separated by a hyper plane.

The limitations of SVM are the selection of a suitable kernel, speed and size, both in training and testing.

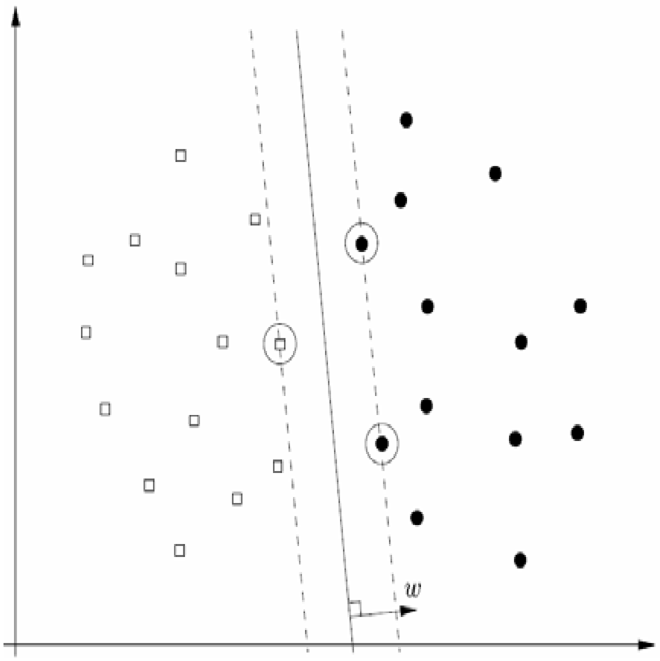


Figure 1: Separating hyperplane for feature selection where circles indicate the support vectors

Multiclass SVM involves the construction of binary SVM classifiers for all pairs of classes. In other words, for every pair of classes, a binary SVM problem is solved (also including the underlying optimization problem to maximize the margin between two classes). An instance is assigned to the class that has the largest number of votes called as Max Wins strategy as per the decision function. If ties still occur, each sample will be assigned a label based on the classification provided by the furthest hyper plane.

One of the benefits of this approach is that for every pair of classes we deal with a much smaller optimization problem. In total we need to solve $k(k-1)/2$ Quadratic Programming (QP) problems which are of size less than *n*. QP optimization algorithms used for SVMs are polynomial to the problem size, hence ample amount of saving can be done in the total computational time.

IV. RESULT

The results of SVM using various kinds of structural and statistical features and their various options are presented in Tables 4 and 6. Table 1 shows the recognition accuracy on different sets of structural features for Bangla script. From Table 4, it can be observed that we get maximum accuracy of 93.27% for Bangla. Similarly, from Table 5 can see that maximum recognition accuracy of 69.67% for Devanagari achieved using statistical features only, when 261 feature vector size consisting of 70 features of zoning (grid size 8), 103 features of Zernike moments (order 17) and 88 features of OFM moments (order 11) have been used.

Table 6 shows the effect of combined statistical features and structural features on the recognition accuracy of degraded printed Bangla as recognition accuracy of 94.78% for Bangla have been observed using a feature vector of size 441. Figure 2 contains the recognition accuracy of different features using SVM classifier.

Table 4: Recognition accuracy of structural features using SVM classifier for Bangla.

Feature	Number of features	Recognition accuracy (%)
St1 to St10	19	26.50
St11	256	86.20
St11	128 (odd)	79.35
St11	128 (even)	81.31
St12	230(M = 5, T = 5)	84.11
St12	165 (M = 5, T = 7)	76.60
St12	115(M =5, T = 10)	66.24
St12	173 (M =4, T = 5)	85.23
St12	127 (M =4, T = 7)	78.41
St12	94 (M = 4, T = 10)	69.80
St12	132 (M = 3, T = 5)	83.37
St12	90 (M = 3, T = 7)	78.70
St12	65 (M = 3, T = 10)	33.14
Combined structural	321 [256 (St11) + 65 (St12)]	93.08
Combined structural	260 [128 (St11) + 132 (St12)]	92.13
Combined structural	193[128 (St11) + 65 (St12)]	93.27
Combined structural	212 [19 (St1 to St10)+ 128 (St11) +65 (St12)]	92.12
Combined structural	407 [19 (St1 to St10)+ 256 (St11) + 132(St12)]	93.18

Table 5: Recognition accuracy of statistical features using SVM classifier for Bangla**.**

Feature	Number of features	Recognition accuracy (%)
Zoning	150 (grid size = 6)	89.30
Zoning	140 (grid size = 7)	86.36
Zoning	70 (grid size = 8)	84.11
Zoning	55 (grid size = 9)	71.10
Zoning	46 (grid size = 10)	68.62
Zoning	33 (grid size = 11)	67.58
Zoning	22 (grid size = 12)	52.33

978-1-4799-8578-4/16/$31.00 ©2016 IEEE

Moments(ZM)	63 (order = 13)	49.76
Moments(ZM)	103 (order = 17)	55.93
Moments(ZM)	160 (order = 21)	59.46
Moments(ZM)	193 (order = 25)	67.51
Moments(OFM)	67 (order = 9)	61.13
Moments(OFM)	88 (order = 11)	65.25
Moments(OFM)	113 (order = 13)	61.37
Combined statistical	388 [140 (zoning)+ 160 (ZM) + 88(OFMM)]	86.13
Combined statistical	318 [70(zoning) +160 (ZM) + 88(OFMM)]	90.07
Combined statistical	303 [55 (zoning) +160 (ZM) + 88(OFMM)]	82.19
Combined statistical	261 [70 (zoning) +103 (ZM) + 88(OFMM)]	90.67
Combined statistical	240 [70 (zoning) +103 (ZM) + 67(OFMM)]	87.67
Combined statistical	221 [70 (zoning) +103 (ZM) + 48(OFMM)]	88.11
Combined statistical	186 [70 (zoning) +68 (ZM) + 48(OFMM)]	87.55

Table 6: Recognition accuracy of all structural and statistical features using SVM classifier for Bangla.

Feature	Number of features	Recognition accuracy (%)
Combined (structural and statistical)	441 [70 (zoning) + 63 (ZM) + 48 (OFMM) + 128 (St11) + 132 (St12)]	94.78
Combined (structural and statistical)	374 [70 (zoning) + 63 (ZM) + 48 (OFMM)+ 128 (St11) + 65(St12)]	94.12

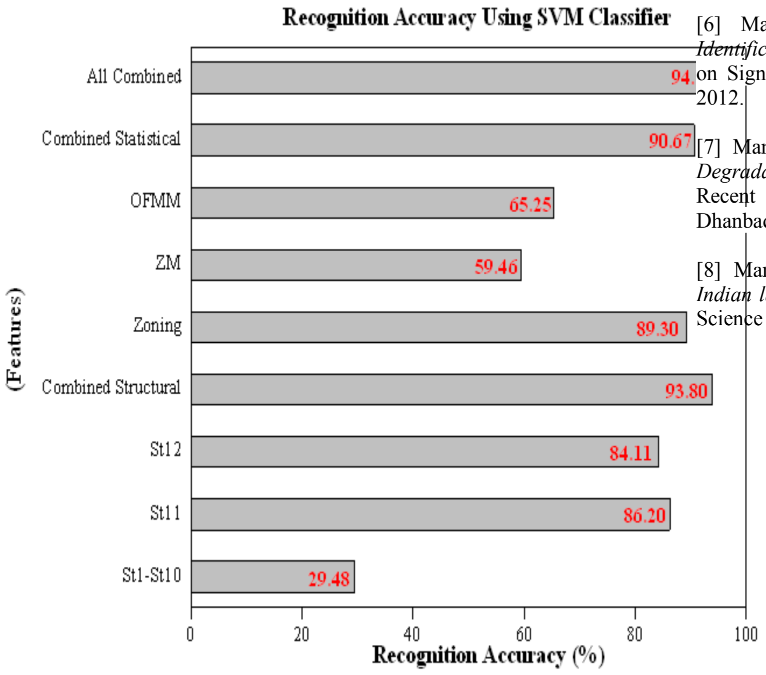


Figure 2: Recognition accuracy for different features using SFAM classifier in Bangla

V. CONCLUSION

This Paper describes a system for script classification of printed Bangla language document. In this Paper we have presented scheme for classification of Bangla script. This paper scheme is based on structural features, statistical features and all combined features of printed Bangla script document. The algorithm used in the paper is very simple, easy to understand and reliable for the classification of the Bangla script document. In the structural features, statistical features and all combined features SVM algorithm, the recognition rate of the scripts classification is very fast and accurate.

REFERENCES

[1] Manoj Kr. Shukla, Haider Banka “Structural Features *Extraction for Devnagari and Bangla language Documents”* Indian Journal of Science & Technology (*ISSN: 0974-6846*), Vol 8(13) July 2015, pp.1-4.

[2] Manoj Kr. Shukla, Haider Banka “*SVM Based classification of the Devnagari Script Document”* MAGNET RESEARCH REPORT *(ISSN: 1444-8939)* , Australia Vol.3 (3) PP: 827-832,

[3] Manoj Kr. Shukla, Haider Banka “*Line-wise Script Segmentation for Indian Language Documents”* International journal of Computer Application” Vol. 108-No.9, ISSN: 0975 - 8887. New York, USA, pp.33-36, 2014.

[4] Manoj Kr. Shukla, Haider Banka “*Degraded Script Identification for Indian Language- A-Survey*” International journal of Computer Application” Vol. 108-No.6,ISSN: 0975 - 8887. New York, USA, pp. 11-22, 2014.

[5] Manoj kr. Shukla, Haider Banka “An *Efficient Segmentation scheme for the Recognition of Printed Devanagari Script.*”. IJCST Vol. 2 Issue 4. 2011 - ISSN: 0976-8491 (online), 2229-4333 (Print).

[6] Manoj Kr. Shukla, Haider Banka, “*Line Based Multi-script Identification for Indian Languages OCR*”,IEEE International Conference on Signal, Image and Video Processing (ICSIVP-2012) , IIT Patna, Jan 2012.

[7] Manoj Kr. Shukla, Haider Banka, “*A Study of Different Kinds of Degradation in Printed Bangla Script*”, IEEE-International Conference on Recent Advances in Information Technology (RAIT-2012) , ISM, Dhanbad, March 2012.

[8] Manoj Kumar Shukla, Manoj Kumar Sharma *“Pre-Processing of Indian languages document images”,* International Journal of Engineering Science (Special issue 2011) – ISSN: 22296913 (Print).