Weka-Based Classification Techniques for Offline Handwritten Gurmukhi Character Recognition

Chapter *in* Advances in Intelligent Systems and Computing · January 2014

DOI: 10.1007/978-81-322-1602-5_76

CITATIONS
4

READS
508

3 authors:

Munish Kumar
Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, INDIA
144 PUBLICATIONS **1,641** CITATIONS
SEE PROFILE

M. K. Jindal
Panjab University
50 PUBLICATIONS **791** CITATIONS
SEE PROFILE

Rishav Kumar Sharma
73 PUBLICATIONS **824** CITATIONS
SEE PROFILE

Weka-Based Classification Techniques for Offline Handwritten Gurmukhi Character Recognition

Munish Kumar, M. K. Jindal and R. K. Sharma

Abstract In this paper, we deal with weka-based classification methods for offline handwritten Gurmukhi character recognition. This paper presents an experimental assessment of the effectiveness of various weka-based classifiers. Here, we have used two efficient feature extraction techniques, namely, parabola curve fitting based features, and power curve fitting based features. For recognition, we have used 18 different classifiers for our experiment. In this work, we have collected 3,500 samples of isolated offline handwritten Gurmukhi characters from 100 different writers. We have taken 60 % data as training data and 40 % data as testing data. This paper presents a novel framework for offline handwritten Gurmukhi character recognition using weka classification methods and provides innovative benchmark for future research. We have achieved a maximum recognition accuracy of about 82.92 % with parabola curve fitting based features and the multilayer perceptron model classifier. In this work, we have used C programming language and weka classification software tool. At this point, we have also reported comparative study weka classification methods for offline handwritten Gurmukhi character recognition.

Keywords Handwritten character recognition · Feature extraction · Classification · Weka · Tool

M. Kumar (✉)
Department of Computer Science, Panjab University Rural Centre,
Kauni, Muktsar, Punjab, India
e-mail: munishcse@gmail.com

M. K. Jindal
Department of Computer Science and Applications, Panjab University Regional Centre,
Muktsar, Punjab, India
e-mail: manishphd@rediffmail.com

R. K. Sharma
School of Mathematics and Computer Applications, Thapar University,
Patiala, Punjab, India
e-mail: rksharma@thapar.edu

B. V. Babu et al. (eds.), *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28–30, 2012*, Advances in Intelligent Systems and Computing 236, DOI: 10.1007/978-81-322-1602-5_76,

1 Introduction

Offline Handwritten Character Recognition usually abbreviated as Offline HCR. Offline HCR is one of the oldest ideas in the history of pattern recognition by using the computer. In character recognition, the process commences by reading of a scanned image of character, determining its meaning, and finally, translates the image into a computer written text document. In recent times, Gurmukhi character recognition has become one of the fields of practical usage. OHCR involves activities like digitization, preprocessing, feature extraction, classification, and recognition. Recognition rate depends on the quality of features extracted from characters and effectiveness of the classifiers. For the past several years, many academic laboratories and companies are occupied with research on handwriting recognition. In the character recognition system, we need three things, i.e., preprocessing on digitized data, feature extraction, and decision-making algorithms. Preprocessing is the initial stage of character recognition. In this phase, the character image is normalized into a window of size 100×100. After normalization, we produce bitmap image of the normalized image. Afterwards, the bitmap image is transformed into a skeletonized image. In this work, we have used two efficient feature extraction techniques, namely, parabola curve fitting based features and power curve fitting based features for character recognition. Aradhya et al. [1] have presented a multilingual OCR system for South Indian scripts based on PCA. Bansal and Sinha [2, 3] have presented a technique for complete Devanagari script recognition. In this technique, they have recognized the character in two steps. In first step, that recognize the unknown characters and in the second step they recognize the character based on the strokes. Chaudhary et al. [4] have represented a technique for recognition of connected handwritten numerals. Gader et al. [6] have presented a handwritten word recognition system using neural network. Hanmandlu et al. [8] have presented a handwritten Hindi numeral recognition system using Fuzzy logic. Kumar [11] has proposed a AI based approach for handwritten Devanagari script recognition. Kumar et al. [12] have presented a review on OCR for handwritten Indian scripts. They have also proposed two efficient feature extraction techniques for offline handwritten Gurmukhi character recognition [13]. Lehal and Singh [14] have presented a printed Gurmukhi script recognition system, where connected components are initially segmented using thinning based approach. Pal et al. [18] have assimilated a comparative study of handwritten Devanagari character recognition using twelve different classifiers and four sets of features. Rajashekaradhya and Ranjan [20] have proposed zoning based feature extraction technique for Kannada script recognition. Roy et al. [21] have presented a script identification system for Persion and Roman script. Sharma et al. [22] have proposed a offline handwritten character recognition system using quadratic classifier. Pal et al. [16, 17] have come up with a technique for offline Bangla handwritten compound characters recognition. They have used modified quadratic discriminant function for feature extraction. They have also presented a technique for feature computation of numeral images. Classification is the most significant activity for character recognition. In the classification process, we required decision making algorithms. There

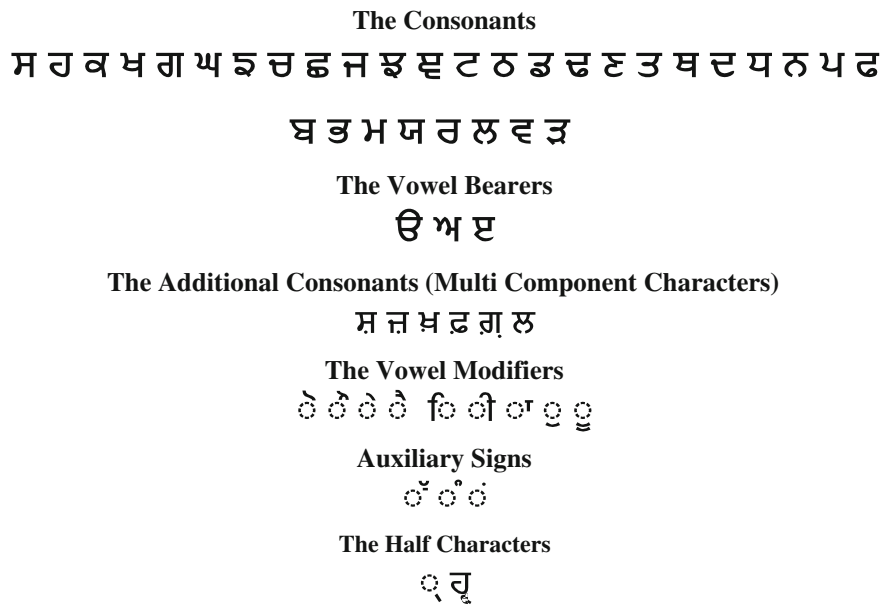

Fig. 1 *Gurmukhi* script character set

have presented various kinds of decision making algorithms as: Baye's Net, DMNB Text, Naïve Baye's, multilayer perceptron model, etc [5, 7, 9, 10, 15, 19]. We have applied 18 different weka classification methods for offline handwritten Gurmukhi character recognition.

2 Gurmukhi Script and Data Collection

Gurmukhi script is the script used for writing in the Punjabi language and is derived from the old *Punjabi* term *Guramukhi*, which means "from the mouth of the Guru". *Gurmukhi* script has three vowel bearers, thirty two consonants, six additional consonants, nine vowel modifiers, three auxiliary signs, and three half characters. The *Gurmukhi* script is the 12th most widely used script in the world. Writing style of *Gurmukhi* script is from top to bottom and left to right. In the *Gurmukhi* script, there is no case sensitivity. The character set of *Gurmukhi* script is given in Fig. 1. In the *Gurmukhi* script, most of the characters have a horizontal line at the upper part called, headline and the characters are connected with one another through this line.

For the present work, we have collected the data from 100 different writers. These writers were requested to writer each Gurmukhi character. A sample of handwritten Gurmukhi characters by five different writers ($W1$, $W2$,…,$W5$) is given in Fig. 2.

Script Character	W_1	W_2	W_3	W_4	W_5
ੳ					
ਅ					
ੲ					
ਸ					
ਹ					

Fig. 2 Samples of handwritten *Gurmukhi* characters

3 Feature Extraction

In this phase, the features of input character are extracted. The performance of Offline HCR system depends on features, which are being extracted. The extracted features should be able to uniquely classify a character. In this work, we have used two efficient feature extraction techniques, namely, parabola curve fitting based features and power curve fitting based features.

3.1 Parabola Curve Fitting Based Features

In this technique, initially, we have divided the thinned image of a character into n (=100) zones. A parabola is fitted to the series of *ON* pixels in every zone by using the least square method. A parabola $y = a + bx + c$ is uniquely defined by three parameters: a, b, and c. This will give $3n$ features for a given bitmap.

The steps that have been used to extract these features are given below.

Step I: Divide the thinned image into n (=100) number of equal sized zones.
Step II: For each zone, fit a parabola using the least square method and calculate the values of a, b and c.
Step III: Corresponding to the zones that do not have a foreground pixel, the values of a, b, and c are taken as zero.

3.2 Power Curve Fitting Based Features

In this technique also, we have divided the thinned image of a character into n (=100) zones. A power curve is fitted to the series of *ON* pixels in every zone, using the least square method. A power curve of the form $y = a$ is uniquely defined by two parameters: a and b. This will give $2n$ features for a given bitmap.

The steps that have been used to extract these features are given below.

Step I: Divide the thinned image into n (=100) number of equal sized zones.
Step II: In each zone, fit a power curve using least square method and calculate the values of a and b.
Step III: Corresponding to the zones that do not have a foreground pixel, the value of a and b is taken as zero.

Table 1 Experimental results of parabola curve fitting based features

Classifier	Accuracy (%)	Root mean squared error (%)	Weighted average precision (%)	False rate (%)	Rejection rate (%)	Weighted F-measure average (%)
Baye's Net	72.78	11.18	74	0.80	26.20	73
Complement naïve bays	57.35	15.61	65.80	1.20	41.40	55.50
DMNB text	72.86	11.58	73.80	0.80	26.25	73
Naïve bays	71.29	12.41	74.30	0.80	27.90	71.70
Multi-layer perceptronl	82.92	8.88	83.40	0.50	16.60	82.90
Multi class classifier	60.86	16.64	65.70	1.10	38	61.80
Classification Via regression	65.57	12.03	66.70	1	33.40	65.60
J48 (decision tree)	55.64	15.11	56.7	1.30	43.10	55.60
IBI	77.85	11.25	79	0.60	20.40	78
IBK	77.86	11.16	80.00	0.60	21.40	78
K-Star	72.07	12.32	74.70	0.80	27.10	72.40
Attribute selected	57.07	14.81	58.80	1.20	41.70	57.30
Ensemble selection	58.07	13.03	59.40	1.20	40.70	57.60
LWL	66.07	12.53	68.70	1	32.90	66.60
PART	53.36	15.66	54.60	1.40	45.20	53.40
Bagging	63.42	12.50	64.60	1.10	35.50	63.30
LogitBoost	62.14	12.11	63.30	1.10	36.80	62.30
Voting feature interval	67.85	12.03	68.10	0.40	31.70	67

Table 2 Experimental results of power curve fitting based features

Classifier	Accuracy (%)	Root mean squared error (%)	Weighted average precision (%)	False rate (%)	Rejection rate (%)	Weighted F-measure average (%)
Baye's Net	72.07	11.50	72.90	0.80	27.10	72.20
Complement naïve bays	61.28	14.87	66.80	1.10	37.60	60.40
DMNB text	72.86	11.58	73.80	0.80	26.30	72.80
Naïve bays	72.86	12.07	75.30	0.80	26.30	73.30
Multi-layer perceptron	82.86	8.80	83.50	0.50	16.60	82.60
Multi class classifier	65.78	16.64	69.30	1.00	33.20	66.60
Classification Via regression	67.85	11.68	68.10	1.00	31.10	67.70
J48 (decision tree)	55.50	15.04	57.10	1.30	43.20	55.60
IBI	77.57	11.32	79.40	0.60	21.80	77.60
IBK	77.57	11.23	79.40	0.60	21.85	77.60
K-Star	69.42	12.74	72.20	0.90	29.70	69.60
Attribute selected	53.14	15.32	54.60	1.40	45.50	53
Ensemble selection	59.42	13.04	60.60	1.20	39.40	59
LWL	64.50	13.05	66.90	1.00	34.50	64.60
PART	55.64	15.18	57.30	1.30	43.10	55.70
Bagging	64.07	12.47	65.10	1.00	34.90	63.80
LogitBoost	64.50	13.05	66.90	1.00	34.50	64.60
Voting feature interval	60.78	16.71	62.60	1.20	38.00	60.00

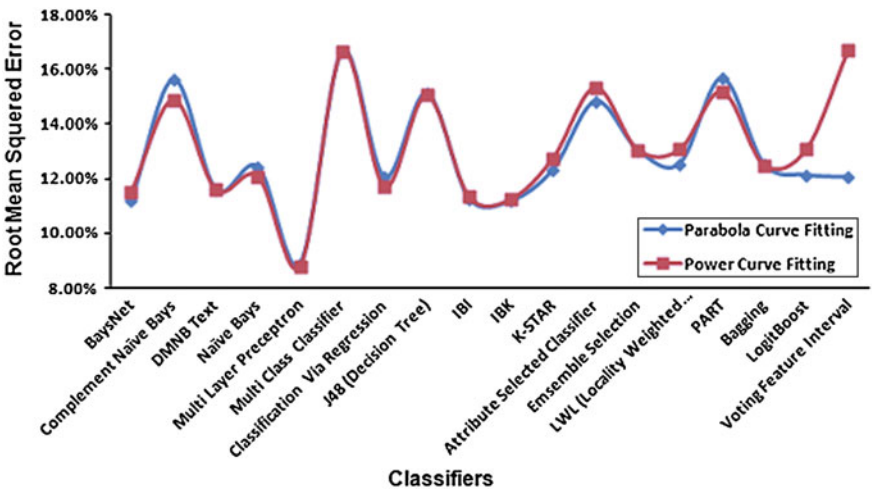

Fig. 3 Recognition accuracy of diverse classification techniques

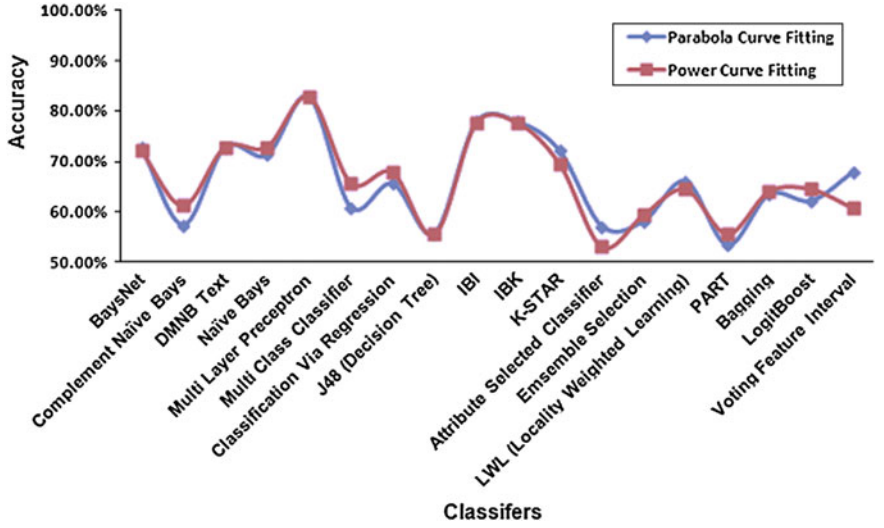

Fig. 4 Comparison of RMSE in various classification techniques

4 Classification

Classification phase is the decision-making phase of an OHCR engine. This phase uses the features extracted in the previous stage for deciding class membership. In this work, we have used 18 different classifiers based on *weka* namely, Baye's Net, Complement Naïve Baye's, Discriminative Multinominal Naïve Baye's Text (DMNB

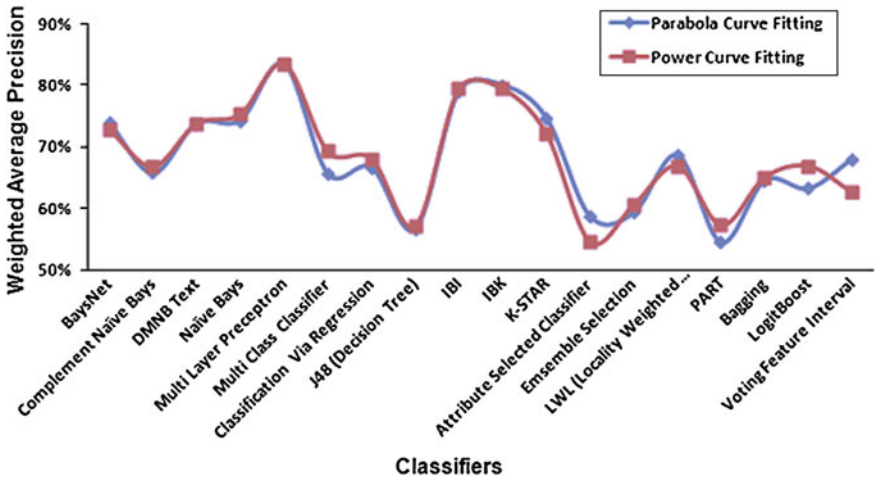

Fig. 5 Weighted average precision of divergent classification techniques

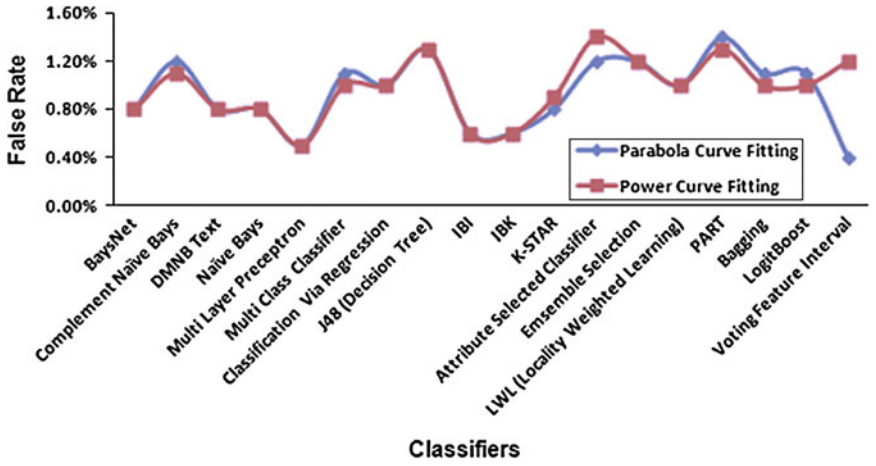

Fig. 6 Comparison of FPRate of different classifiers

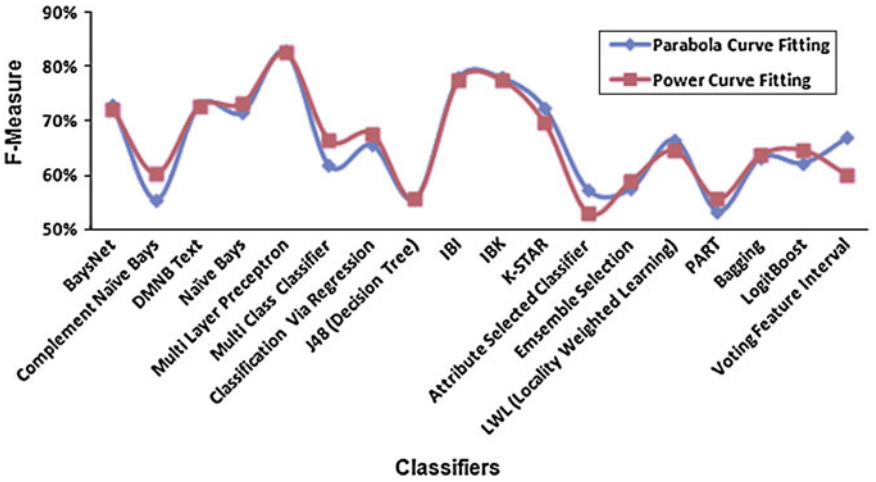

Fig. 7 F-Measure of diverse classification methods

Text), Naïve Baye's, Multi-Layer Perceptron, Multi Class Classifier, Classification via Regression, Decision Tree, IBI, IBK, K-Star, Attribute Selected Classifier, Ensemble Selection, Locality Weighted Learning (LWL), PART, Bagging, LogitBoost, and Voting Feature Interval classifier for offline handwritten *Gurmukhi* character recognition. Experimental results of these classifiers are presented in the next section.

5 Experimental Results

In this section, we have presented results of various *weka*-based classification methods. Table 1, depicts the results based on parabola curve fitting based feature extraction technique and Table 2 shows the power curve fitting based feature extraction results. In Fig. 3, we have presented recognition accuracy of various classification techniques for offline handwritten *Gurmukhi* character recognition. As such, we have seen that the multilayer perceptron model is the preeminent classifier for offline handwritten character recognition. We have achieved a maximum accuracy of about 82.92 % with parabola curve fitting based features and the multilayer perceptron model classifier.

In Fig. 4, we have presented the root mean squared error (RMSE) of various classification techniques, based on parabola curve fitting features and power curve fitting features. Figure 5, signifies the precision of different classification techniques. In Fig. 6, false rate (FP) of different classification techniques is depicted, graphically. Figure 7, describes the F-measure of various classifiers, considered in this work.

6 Conclusion

In present work, we have illustrated the effectiveness of various *weka*-based classification techniques for offline handwritten *Gurmukhi* character recognition. We have experimented on our own data set. We have collected 3,500 samples of isolated handwritten Gurmukhi characters from 100 different writers. After making a comparison of various classifiers for the recognition of character, we drew conclusion that the most appropriate classification technique is multilayer perceptron model. We obtained a maximum accuracy of about 82.92 % with parabola curve fitting based features and multilayer perceptron model classifier. The results of power curve fitting are also promising. A further study of the benefits of this technique can also be extended to offline handwritten character recognition of other Indian scripts.

References

1. Aradhya, V.N.M., Kumar, G.H., Noushath, S.: Multilingual OCR system for south Indian scripts and English documents: an approach based on Fourier transform and principal component analysis. Eng. Appl. Artif. Intell. **21**, 658–668 (2008)
2. Bansal, V., Sinha, R.M.K.: Integrating knowledge sources in devanagari text recognition. IEEE Trans. Syst. Man Cybern. **30**(4), 500–505 (2000)
3. Bansal, V., Sinha, R.M.K.: Segmentation of touching and fused Devanagri characters. Pattern Recogn. **35**(4), 875–893 (2002)
4. Chaudhuri, B.B., Majumder, D.D., Parui, S.K.: A procedure for recognition of connected hand written numerals. Int. J. Syst. Sci. **13**, 1019–1029 (1982)
5. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Mach. Learn. **29**, 131–163 (1997)

6. Gader, P.D., Mohamed, M., Chaing, J.H.: Handwritten word recognition with character and inter-character neural networks. IEEE Trans. Syst. Man Cybern. Part B Cybern. **27**(1), 158–164 (1997)
7. Genkin, A., Lewis, D., Madigan, D.: Large-scale Bayesian logistic regression for text categorization. TECHNOME TRICS. **49**(3), 291–304 (2004)
8. Hanmandlu, M., Grover, J., Madasu, V. K. and Vasikarla, S.: Input fuzzy for the recognition of handwritten Hindi numeral. In: Proceedings of ITNG, pp. 208–213 (2007)
9. Jomy, J., Parmod, K.V., Kannan, B.: Handwritten character recognition of south Indian scripts: a review. In: Proceedings of Indian Language Computing, pp. 1–6 (2011)
10. Kala, R., Vazirani, H., Shukla, A. and Tiwari, R.: Offline handwriting recognition using genetic algorithm. Int. J. Comput. Sci. Issues **7**(2,1), 16–25 (2010)
11. Kumar, D.: AI approach to hand written Devanagri script recognition. In: Proceedings of 10th International conference on EC3-Energy, Computer, Communication and Control Systems, **2**, 229–237 (2008)
12. Kumar, M., Jindal, M. K. and Sharma, R. K.: Review on OCR for handwritten Indian scripts character recognition. In: Proceedings of DPPR, pp. 268–276 (2011)
13. Kumar, M., Jindal, M.K., Sharma, R.K.: Efficient feature extraction technique for offline handwritten Gurmukhi character recognition. Chaing Mai J. Sci. (2012)
14. Lehal, G.S., Singh, C.: A Gurmukhi script recognition system. In: Proceedings of 15th ICPR, pp. 557–560 (2000)
15. McCallum, A., Nigam, K.: A comparison of event models for Naive Baye's text classification. Paper presented at the workshop on learning for text categorization (1998)
16. Pal, U., Wakabayashi, T., Kimura, F.: Handwritten numeral recognition of six popular scripts. In: Proceedings of ICDAR, vol. **2**, pp. 749–753 (2007a)
17. Pal, U., Wakabayashi, T., Kimura, F.: A system for off-line Oriya handwritten character recognition using curvature feature. In: Proceedings of 10th ICIT, pp. 227–229 (2007b)
18. Pal, U., Wakabayashi, T., Kimura, F.: Comparative study of devanagari handwritten character recognition using different feature and classifiers. In: Proceedings of 10th ICDAR, pp. 1111–1115 (2009)
19. Patel, C.I., Patel, R., Patel, P.: Handwritten character recognition using neural network. Int. J. Sci. Eng. Res. **2**(4), 1–5 (2011)
20. Rajashekararadhya, S.V., Ranjan, S. V.: Zone based feature extraction algorithm for handwritten numeral recognition of Kannada script. In: Proceedings of IACC, pp. 525–528 (2009)
21. Roy, K., Alaei, A., Pal, U.: Word-wise handwritten Persian and Roman script identification. In: Proceedings of ICFHR, pp. 628–633 (2010)
22. Sharma, N., Pal, U., Kimura, F., Pal, S.: Recognition of off-line handwritten devanagari characters using quadratic classifier. In: Proceedings of ICVGIP, pp. 805–816 (2006)