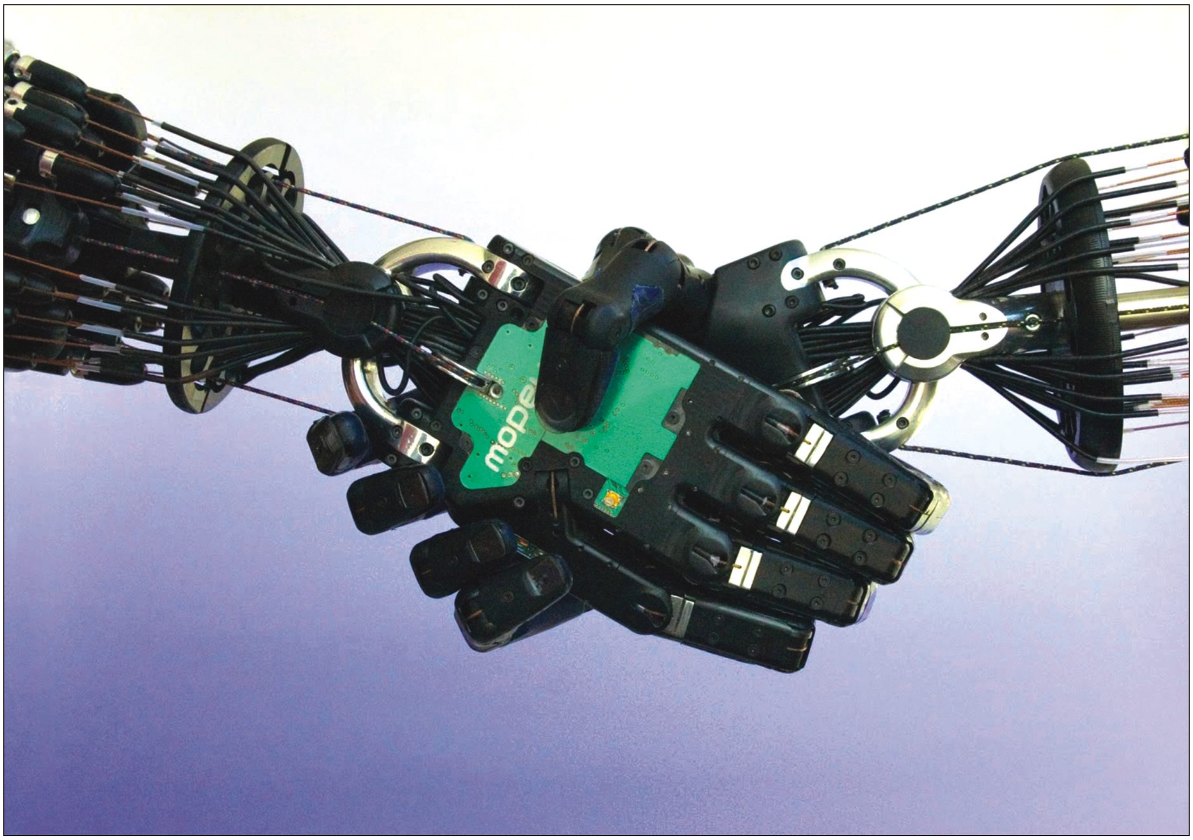

Journal of Artificial Intelligence

ISSN 1994-5450

OPEN ACCESS Journal of Artificial Intelligence

ISSN 1994-5450
DOI: 10.3923/jai.2018.65.70

Mini Review
Devanagari and Gurmukhi Script Recognition in the Context of Machine Learning Classifiers

Reya Sharma, Baij Nath Kaushik and Naveen Kumar Gondhi

School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra-182320, J and K, India

Abstract
The handwritten character recognition is potentially an active area of research due to the presence of several challenging issues. Due to a large variation in writing styles, development of optical handwritten character reader is a challenging task. In order to decrease the burden of computation and to improve the recognition accuracy, several measures need to be taken in the overall process of recognition. The main objective of this review was to recognize and analyze handwritten document images. There are wide varieties of classification techniques available for the problem of pattern recognition. These techniques include Support Vector Machine (SVM), Back Propagation Neural Networks (BPNN), Probabilistic Neural Networks (PNN) and many more. In this study, a survey has been performed on some of these machine learning techniques for the identification of various handwritten north Indian scripts. This study attempts to address the most significant results obtained so far and then all the gathered data is represented in the form of tables so as to have a clear idea by visualizing data at once. This research paper provides a comprehensive survey on various machine learning techniques involved in north Indian script recognition and the study also highlights the crucial aspects of the research till date.

Key words: Handwritten character recognition, optical character recognition (OCR), machine learning, classification, devanagari, gurmukhi

Citation: Reya Sharma, Baij Nath Kaushik and Naveen Kumar Gondhi, 2018. Devanagari and gurmukhi script recognition in the context of machine learning classifiers. J. Artif. Intel., 11: 65-70.

Corresponding Author: Baij Nath Kaushik, School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra-182320, J and K, India Tel: 9654482709

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

The optical character recognition (OCR) is a technique of mechanical or electronic transformation of images of printed text or handwritten text into the machine-readable form. The OCR is an emerging field of research in artificial intelligence, machine vision and pattern recognition. Now-a-days the recognition accuracy of printed text is considered potentially a solved problem. However, due to the large variation in writing styles the handwritten text recognition is considered to be a difficult task[1]. Hence, the handwritten character recognition is potentially an active field of research. The OCR works by implementing several techniques and tools and based upon them the classification accuracy is determined.

Translation of documents from scanned form to a machine-readable form leads us towards the paperless environment which indicates the idea of optical character recognition system. There is a high demand of OCR for several emerging applications such as postal systems, word processing, banks, etc. The main objective of OCR is to recognize and analyze an image document with the help of splitting the document into lines and then further splitting these lines into words and finally into characters[2]. From these characters, features are extracted and then these features are compared with the image patterns in order to identify characters. The identification of the characters can be carried out either from handwritten documents or printed documents. The recognition of handwritten document can be carried out either offline or online. In case of online identification temporal information is available like coordinates of pens tip as a function of time, however, in case of offline identification, the image of the handwritten document is available only for the computer[3].

There is a wide variety of classification techniques available for the problem of pattern recognition. These techniques include several methods like Support Vector Machine (SVM), Back Propagation Neural Networks (BPNN), Extreme Learning Machine (ELM), Probabilistic Neural Networks (PNN), Hidden Markov Model, Convolutional Neural Networks (CNN), K-Nearest Neighbors (KNN), etc. In this paper, some of these machine learning techniques are discussed in detail for the identification of handwritten characters. Most of the classification strategies in optical character reader deal with a potentially huge number of classes aiming towards finding discrimination among them.

INTRODUCTION TO SCRIPTS

The majority of languages spoken in India are the Indo-aryan languages. Most of the languages spoken in India are inferred from the ancient Brahmi script are Devanagari, Gurmukhi, Urdu, Kannada, Bangla, Telugu, Malayalam, Gujarati, Tamil and Oriya[4]. Among them, the most commonly used north Indian scripts are Devanagari and Gurmukhi.

Devanagari script is one among the widely used and adopted system for writing in India. More than 500 million people use Devanagari script and almost 120 languages are derived from this script like Sanskrit, Rajasthani, Hindi, Haryanvi, Dogri, Kashmiri, Bhojpuri, Marathi and much more[5]. Devanagari basically consists of 47 characters and from these characters 14 are vowels and remaining 33 are consonants. The direction of the writing system for Devanagari script is from left to right.

Gurmukhi is another common script of north India and mainly used in Punjab. Over 29 million people use Gurmukhi script and it is used for writing the Punjabi language[6]. Its writing system follows the direction from left to right. Gurmukhi basically consists of 44 primary characters and from these characters 9 are vowels and remaining 35 are consonants. Guru Granth Sahib which is the prime scripture of Sikhism is written using Gurmukhi script.

The classification aims towards reducing the possible number of characters for an unidentified character from the know characters. The appropriate selection of classifier is not at all an easy task because there are various factors on which the classifier depends, for example, number of free parameters, the available training set and much more[7]. For the recognition of characters, various classification methods that are based on learning are used. Some of these classifiers are like Artificial Neural Network[8-11], Probabilistic Neural Network, Support Vector Machine, Hidden Markov Model[12], K-Nearest Neighbors, etc. For such classification techniques patterns are fed during the training phase and the system regulates itself so as to decrease misclassification error for such patterns[13].

Probabilistic neural network classifier: Probabilistic Neural Network is a suitable form of network for the classification problems. Donald Specht proposed the Probabilistic Neural Network. The PNN classifier has fast learning ability and it also provides a good generalization capability which is hugely important for recognition of handwritten characters[14]. The PNN is established on the basis of concepts used in traditional pattern recognition problems. The PNN models are generally based on a well-known technique known as the Bayesian classifier which diminishes the probable risk of recognizing patterns in false category[15]. The PNN makes use of non-parametric technique which uses parzen window and this parzen window creates a class of estimates from the kernels. Equation 1 gives the classical form of this estimator:

$$\frac{1}{n\sigma}\sum w((z-z_i)|\sigma) \quad (1)$$

where, σ is the smoothing parameter, w indicates the weighting function, z is the unknown input and z_i is the ith sample.

Generally, the Gaussian function is applied as the weighting function. The advantage of using PNN for classification problem is its fast ability to be trained as well as it can work with data having points beyond the norms and hence has better performance than several other neural network architectures.

Support vector machine: Support Vector Machine classifier is basically modeled on the statistical learning techniques. The SVM is a very strong classifier and used largely with success for the pattern recognition and classification tasks. SVM was introduced by Surinta *et al.*[16]. The SVM was basically used for two class problem in which it gazes for optimal hyperplanes which increases the margin and distance between two classes. In order to determine the hyperplane normal vector is used, which can be represented by linear combination of closest examples of the two classes called as support vectors. Kernels techniques can be used so as to extend SVM for more conventional decision surfaces rather than hyperplanes. SVM is regarded to be state-of-art for solving the nonlinear and linear classification problems. The advantages of SVM are its flexibility, the global optimum character, prediction capacity and parsimoniousness[17]. The formulation of SVM is based upon structural risk minimization which is considered to be better than empirical risk minimization used commonly in the artificial neural networks.

Convolutional neural networks: The convolutional neural networks based on a differentiable function which transforms 3-dimensional volume input into 3-dimensional volume output and it has multiple hidden layers[18]. These hidden layers can be pooling, convolutional or fully connected. In order to obtain a class score, CNN transforms the actual images through every layer in the network. As the input of CNN consists of images, therefore, the neurons in the hidden convolution layer possess an activation volume with height, depth and width dimensions. In order to improve the representative power of the architecture, the activation layers use element-wise operations so as to bring non-linearity. In order to handle over fitting pooling layers are introduced so as to decrease the spatial size of representation and thus decreasing the number of parameters. Additional regularization is provided by drop out layers only by holding each neuron active[19].

The last layer of convolutional neural networks represents a soft-max classifier and where $f(x_k, W) = Wx_k$ is the mapping function which provides scores, interpreted for each class as the un-normalized log probabilities. Class scores are calculated with the help of using soft-max function as given below in Eq. 2:

$$f_i(z) = \frac{e^{zi}}{\sum e^{zj}} \quad (2)$$

where, z indicates the score vector and f_i indicates the vector having values in between 0 and 1 that results in sums to 1. Equation 3 computes the corresponding loss function in which f_i represents the ith element of f, the class score vector.

$$L_k = -\log\frac{e^{fyk}}{\sum e^{fi}} \quad (3)$$

Extreme learning machine: Extreme learning machine is basically a single layer feed-forward neural network with the hidden layer of neurons[20]. The ELM is a very strong technique for pattern recognition and classification due to its several advantages like fast learning speed as compared to the classical gradient-based learning techniques[21]. ELM has several other advantages like easy implementation, avoid local optimal value and better generalization performance. These all because of the fact that the parameters in ELM such as biases and input weights are assigned randomly i.e., the output weights are computed analytically to fine tune the ELM algorithm. The ELM has thousand times better learning speed than that of classical feed-forward neural algorithms such as back-propagation.

Work already done on Devanagari and Gurmukhi scripts by analyzing various classification and feature extraction techniques in terms of the accuracy provided by them.

Table 1: Details of work done on handwritten Devanagari script

References	Data type	Feature	Classifier	Dataset size	Accuracy (%)
Pal and Sontakke[22]	Numerical	Gradient	Modified quadratic discriminant function	22556	99.56
Patil *et al.*[23]	Numerical	Structural	Fuzzy neural network	2000	99.50
Bhattacharya and Chaudhuri[24]	Numerical	Wavelet	Multilayer perceptron	22556	99.27
Sharma *et al.*[25]	Numerical	Chaincode	Quadratic	22556	98.86
Jangid *et al.*[26]	Numerical	Zoning, background directional distribution	Support vector machine	22556	98.62
Bhattacharya *et al.*[27]	Numerical	Scalar features	Hidden markov model	22535	87.69
Pal *et al.*[28]	Character	Gradient	Mirror image learning	36172	95.19
Pal *et al.*[29]	Character	Gradient	Support vector machine	36172	95.13
Mane and Ragha[30]	Character	Eigen deformation	Elastic matching	3600	94.91
Pal *et al.*[31]	Character	Gradient	Quadratic	36172	94.24

Table 2: Details of work done on handwritten Gurmukhi script

References	Data type	Feature	Classifier	Dataset size	Accuracy (%)
Sinha *et al.*[32]	Numerical	Zoning	Support vector machine, radial basis function kernel	1500	99.73
Garg[33]	Character	Structural	Neural network	6900	96.00
Kartar *et al.*[34]	Character	Zoning, background directional distribution	Support vector machine	7000	95.04
Sharma and Jhajj[35]	Character	Zoning	Support vector machine, K-Nearest neighbour	5125	73.02

Devanagari: Structural features generally provide better results for the handwritten symbol recognition. Table 1 indicated the summarized results obtained so far in handwritten Devanagari character recognition. As shown in Table 1, the maximum accuracy obtained in case of handwritten Devanagari numerical recognition[22] is 95.56% and the maximum accuracy obtained in case of handwritten Devanagari character recognition[28] is 95.19%. Various classifiers used for the classification of handwritten Devanagari script are like SVM, MIL, MQFD, Quadratic, Fuzzy Neural Network, Elastic Matching, HMM, etc. Different feature extraction techniques used for recognition of handwritten Devanagari script are Gradient, Structural, Chain Code, Wavelet, Zoning density, Scalar features, etc.

Gurmukhi: Data in Table 2 indicated the summarized results obtained so far in handwritten Gurmukhi symbol recognition. As shown in Table 2, the maximum accuracy obtained in case of handwritten Gurmukhi numerical recognition[32] is 99.73% and the maximum accuracy obtained in case of handwritten Gurmukhi character recognition[33] is 96%. Various classifiers used for the classification of handwritten Gurmukhi script are like SVM with RBF Kernels, Neural Network, KNN, etc. Different feature extraction techniques used for recognition of handwritten Gurumukhi script are Structural, Zone based, Zoning density, etc.

CONCLUSION

This study concluded the work done on various machine learning techniques for the identification of various handwritten north Indian scripts like Devanagari and Gurmukhi, then all the gathered data is represented in the form of tables in order to provide a clear idea by visualizing data at once. This study also discusses in detail various classifiers like SVM, PNN, CNN and ELM used for the classification and recognition of handwritten characters. This study point outs that the work done on north Indian scripts is still at its infancy, so this area still needs further attention to resolve many problems.

SIGNIFICANCE STATEMENT

Although in the last few years there has been significant advances done in the field of script recognition, however a huge amount of work is still needed to be done in order to improve the efficiency and accuracy of script recognition systems. Some most beneficial outlooks of research are addressed in the study. Furthermore, this study also points out that there is a lack of availability of standardized data set for the handwritten north Indian scripts. In future, researchers can put significant efforts in order to build benchmark databases for handwritten North Indian scripts so as to improve the quality of study related to OCR. Since the study done on the handwritten Devanagari and Gurmukhi script is limited, so there is a wide scope for more research in these areas of handwritten script recognition for the future researchers.

REFERENCES

1. Bhunia, A.K., P.P. Roy, A. Mohta and U. Pal, 2017. Cross-language framework for word recognition and spotting of indic scripts. Pattern Recogn., 79: 12-31.

2. Jindal, K. and R. Kumar, 2018. A new method for segmentation of pre-detected Devanagari words from the scene images: Pihu method. Comput. Electr. Eng., 70: 754-763.
3. Singh, P.K., R. Sarkar and M. Nasipuri, 2015. Offline script identification from multilingual Indic-script documents: A state-of-the-art. Comput. Sci. Rev., 15: 1-28.
4. Ubul, K., G. Tursun, A. Aysa, D. Impedovo, G. Pirlo and T. Yibulayin, 2017. Script identification of multi-script documents: A survey. IEEE Access, 5: 6546-6559.
5. Khanale, P.B. and S.D. Chitnis, 2011. Handwritten devanagari character recognition using artificial neural network. J. Artif. Intell., 4: 55-62.
6. Aggarwal, A., K. Singh and K. Singh, 2015. Use of gradient technique for extracting features from handwritten gurmukhi characters and numerals. Proc. Comput. Sci., 46: 1716-1723.
7. Agarwal, M. and B. Kaushik, 2016. Review of text recognition works. Int. J. Comput. Applic., 137: 34-39.
8. Kaushik, B., N. Kaur and A.K. Kohli, 2013. Achieving maximum reliability in fault tolerant network design for variable networks. Applied Soft Comput., 13: 3211-3224.
9. Khanale, P.B., 2010. Recognition of marathi numerals using arti cial neural network. J. Artif. Intell., 3: 135-140.
10. Kaushik, B. and H. Banka, 2015. Performance evaluation of Approximated Artificial Neural Network (AANN) algorithm for reliability improvement. Applied Soft Comput., 26: 303-314.
11. Kaushik, B. and H. Banka, 2014. Approach for improving reliability in optimal network design. Int. J. Adv. Intell. Paradigms, 6: 157-175.
12. Roy, P.P., A.K. Bhunia, A. Das, P. Dey and U. Pal, 2016. HMM-based Indic handwritten word recognition using zone segmentation. Pattern Recogn., 60: 1057-1075.
13. Agarwal, M. and B. Kaushik, 2015. Text recognition from image using artificial neural network and genetic algorithm. Proceedings of the International Conference on Green Computing and Internet of Things, October 8-10, 2015, Noida, India, pp: 1610-1617.
14. Lotfi, A. and A. Benyettou, 2011. Using probabilistic neural networks for handwritten digit recognition. J. Artif. Intell., 4: 288-294.
15. Dhandra, B.V., R.G. Benne and M. Hangarge, 2010. Kannada, Telugu and Devanagari handwritten numeral recognition with probabilistic neural network: A novel approach. Int. J. Comput. Applic., 26: 83-88.
16. Surinta, O., M.F. Karaaba, L.R.B. Schomaker and M.A. Wiering, 2015. Recognition of handwritten characters using local gradient feature descriptors. Eng. Applic. Artif. Intell., 45: 405-414.
17. Sahare, P. and S.B. Dhok, 2018. Multilingual character segmentation and recognition schemes for Indian document images. IEEE Access, 6: 10603-10617.
18. Sarkhel, R., N. Das, A. Das, M. Kundu and M. Nasipuri, 2017. A multi-scale deep quad tree based feature extraction method for the recognition of isolated handwritten characters of popular indic scripts. Pattern Recogn., 71: 78-93.
19. Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res., 15: 1929-1958.
20. Huang, G.B., Q.Y. Zhu and C.K. Siew, 2004. Extreme learning machine: A new learning scheme of feedforward neural networks. Proceedings of the IEEE International Joint Conference on Neural Networks, July 25-29, 2004, Budapest, Hungary, pp: 985-990.
21. Zhu, Q.Y., A.K. Qin, P.N. Suganthan and G.B. Huang, 2005. Evolutionary extreme learning machine. Pattern Recognition, 38: 1759-1763.
22. Pal, U., N. Sharma, T. Wakabayashi and F. Kimura, 2007. Handwritten numeral recognition of six popular Indian scripts. Proceedings of the 9th International Conference on Document Analysis and Recognition, September 23-36, 2007, Parana, Brazil, pp: 749-753.
23. Patil, P.M. and T.R. Sontakke, 2007. Rotation, scale and translation invariant handwritten Devanagari numeral character recognition using general fuzzy neural network. Pattern Recogn., 40: 2110-2117.
24. Bhattacharya, U. and B.B. Chaudhuri, 2009. Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals. IEEE Trans. Pattern Anal. Mach. Intell., 31: 444-457.
25. Sharma, N., U. Pal, F. Kimura and S. Pal, 2006. Recognition of Off-Line Handwritten Devnagari Characters using Quadratic Classifier. In: Computer Vision, Graphics and Image Processing, Kalra, P.K. and S. Peleg (Eds.). Springer, Berlin, ISBN: 978-3-540-68302-5, pp: 805-816.
26. Jangid, M., R. Dhir, R. Rani and K. Singh, 2011. SVM classifier for recognition of handwritten devanagari numeral. Proceedings of the International Conference on Image Information Processing, November 3-5, 2011, Shimla, India, pp: 1-5.
27. Bhattacharya, U., S.K. Parui, B. Shaw and K. Bhattacharya, 2006. Neural combination of ann and hmm for handwritten devanagari numeral recognition. Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition, October 2006, La Baule, France.
28. Pal, U., T. Wakabayashi and F. Kimura, 2009. Comparative study of devnagari handwritten character recognition using different feature and classifiers. Proceedings of the 10th International Conference on Document Analysis and Recognition, July 26-29, 2009, Barcelona, Spain, pp: 1111-1115.

29. Pal, U., S. Chanda, T. Wakabayashi and F. Kimura, 2008. Accuracy improvement of devnagari character recognition combining SVM and MQDF. Proceedings of the 11th International Conference on Frontiers Handwriting Recognition, August 19-21, 2008, Montreal, Quebec, Canada, pp: 367-372.
30. Mane, V. and L. Ragha, 2009. Handwritten character recognition using elastic matching and PCA. Proceedings of the International Conference on Advances in Computing, Communication and Control, January 23-24, 2009, Mumbai, India, pp: 410-415.
31. Pal, U., N. Sharma, T. Wakabayashi and F. Kimura, 2007. Off- line handwritten character recognition of devnagari script. Proceedings of the 9th International Conference on Document Analysis and Recognition, September 23-26, 2007, Parana, Brazil, pp: 496-500.
32. Sinha, G., R. Rani and R. Dhir, 2012. Handwritten gurmukhi numeral recognition using zone-based hybrid feature extraction techniques. Int. J. Comput. Applic., 47: 24-29.
33. Garg, N., 2009. Handwritten gurumukhi character recognition using neural networks. Master Thesis, Computer Science and Engineering Department Thapar University, Patiala, India.
34. Siddharth, K.S., R. Dhir and R. Rani, 2011. Handwritten gurumukhi character recognition using zoning density and background directional distribution features. Int. J. Comput. Sci. Inform. Technol., 2: 1036-1041.
35. Sharma, D. and P. Jhajj, 2010. Recognition of isolated handwritten characters in gurmukhi script. Int. J. Comput. Applic., 4: 9-17.