International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

An Insight of Script Text Extraction Performance using Machine Learning Techniques

Shikha Chadha , Sonu Mittal ,Vivek Singhal

Abstract: ***With the evolution of huge amount of ancient and modern text available in digital format, it is ascertain to mine for researchers, government, tourist and travelers visiting all over the world. However, it is very challenging and costly. Further, it takes a lot of effort and time for script text mining. Therefore, the study investigates various techniques for script text mining viz supervised and unsupervised techniques. Firstly, the study presents a survey for various kinds of techniques adopted by the users for extraction of text from image. It also delivers information about gaps involved in the various approaches. Furthermore, it incorporate the quantitative comparisons based among the study of various approaches and techniques for text extraction as well as script level comparison. The result inferred on the basis of the script comparison indicates that, the accuracy level of ancient script was found to be 5% lesser than modern script. Again, furthermore comparison has been done on standalone and hybrid machine (Combination of CNN and KNN) / deep learning techniques. The accuracy has been found to be lower(4%) in case of standalone techniques.***

KEYWORDS: ***Text Mining, Machine Learning, Deep Learning, Script, Image scene extraction, Convolutional Neural Network, K-Nearest Neighbors.***

I. INTRODUCTION

Text mining and analysis involves the conversion of unstructured data into a structured form for analysis, visualization, etc. The text mining is an collaborative approach, which embosses the Information and communication Technology (ICT) [28] viz. extraction of data , mining of data using and its computational statistics [5] and Internet of things (IOT) [45]. It encompasses the significant areas in Natural Language Processing (NLP) [37], which proved as an ice breaker for different people communicating in various languages, in order to communicate among each other to share ideas, culture, etc [34].

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Shikha Chadha*, Ph.D. Research Scholar, Jaipur National University,Rajasthan, India. shikha1232@gmail.com.
Sonu Mittal, Associate Professor Department of Computer and System Sciences, Jaipur National University, Rajasthan, India. dr.sonumittal@jnujaipur.ac.in
Vivek Singhal, Associate Professor, Department of Information &Technology, JSSATE, Noida, (U.P), India . vivek.singhal@jssaten.ac.in

World is very rich in ancient culture, scientific knowledge flowing in form of Vedas, Upanishads which is a course of attraction for foreigners to explore the rich culture of the world and as most of the scripts are available in degraded form [20],[42]. Text mining involves various steps; firstly text preprocessing is carried out, which involves the digitization of handwritten carved manuscripts as input text. Secondly, after pre-processing the input image is free from noise or disturbance and is carried to the next phase i.e segmentation, in which the image is broken into discrete characters, lines, and paragraphs. Further, text feature extraction involves the extracting the relevant information from the raw data to be categorized as global or statistical features. Once the text extraction is performed, next step in text mining is machine translation which incorporates conversion of source language to target languages using various approaches is done using conventional machine learning techniques. Mostly the exhaustive challenges faced in text mining are the text extraction. Segregating the text and non-text area from the scene text, containing a lot of disturbance and in-text translation i.e. understanding the semantic meaning of the text and thereafter, it is translated into equivalent target text [24],[11].

As a result, a lot of study has been engaviesd to speculate new studies since many years. The focus of the study is to accomplish the most efficient challenges faced for text mining and evaluate the most efficient techniques for ancient script extraction using machine learning techniques.

Techniques used for Text Mining

A. Supervised Learning

Supervised learning is machine learning which infers the classifier from training data set to extract the knowledge from the given dataset. Various supervised learning algorithms like K-nearest neighbor, decision tree, rule-based, etc. are used for text mining [16].

B. Unsupervised Learning Methods

Unsupervised learning method is a technique of extracting information from the unlabeled text with no training data available .This technique can be applied on any text data. Clustering and text modeling are two approaches used for unsupervised learning methods [23] .It is the process of grouping similar text together by finding the similarities between the various texts and segmenting them.

Text modeling is a probabilistic model used to determine a soft clustering, in which every text has a probability detection [36] over all the clusters.

C. Recurrent neural network (RNN)

RNN is a kind of neural network in which output extracted from the previous layer is given as an input to the next layer [21], as in neural network input and output is independent of each other but sometimes for predicting next word information about previous word is required .Therefore, RNN came into light which solved the problem with hidden layer which is required for managing the sequential data. The RNN aims to anticipate the tag of current layer using provisional information of previous layer. It generally requires an exhaustive training process as the error decreases exponentially along the sequence [36] due to which being an dominant classification model it is still not used for literature.

D. Convolutional Neural Networks (CNN)

CNN is a technique of deep learning that accepts input image and assign weights and biases to various objects in the image to differentiate one object from another. The main advantage of CNN is very less than other preprocessing algorithms [9],[33]. In order to digitize the text, it performs Optical Character Recognition (OCR) and converts it into a convenient format for Natural Language Processing (NLP) of handwritten or printed text [21].

E. *Convolutional Recurrent Neural Network (CRNN)*

It is the combines the illustrative features of CNN and RNN, whereas CNN is used for feature extraction and RNN is used for encoding and decoding of feature sequence extracted. [15],43].

II. QUANTATIVE RESEARCH

The study is based on the two comparison between various approaches used for script text extraction among twenty-nine studies to choose the challenge in each approach and is show it's Table.1 and also shows effect on detection accuracy [39]. Firstly comparison has been done based on various machine learning techniques used for text extraction. Second comparison is done on various deep learning techniques and Image Processing techniques for text image extraction in Table 1 demonstrates the contrast between the twenty nine researches about script text extraction and research gaps. The study analysis results in various essential factors important and relevant in review structure and domain for each review structure. Another comparison has been done based on input and output parameters in various research papers as given in Table 1 also provides the proposed framework for various text extraction and noise removal techniques when dataset provided is in degraded format, which is another major challenge for researcher to extract the text.

Table-I: Study for the Text Extraction Challenges

S.No	Author	Domain	Input	Output	Review	Gaps
1	A.Jeyapriya & C.S.Kanimozhi Selvi	Machine Learning	Online Product Review	Opinion Orientation	Framework for extracting the opinion for the product review	To summarize the aspects based on the relative importance of the extracted aspect.
2	Bijalwan et al.(2014)	Machine Learning	News articles	Vector Space Model for Information Retrieval application Using	proposed a supervised machine learning technique using KNN for text mining .	Time complexity is high
3	Trstenjak.B (2014)	Machine Learning	Document from Online Sources of different sizes and categories are used	Weight Matrix	Framework for text classification containing five modules GUI, Preprocessing, KNN& Tf-idf, measuring module and document module	More amount and category of data to be used
4	Gopal&Raghav(2017)	Machine Learning		Term Document Matrix	Classifier based Automatic document retrieval system	System applies to English characters , it can be extend for different languages like Hindi, Kannada, etc. and handwritten characters.
5	Niusha Shafiabady (2015)	Machine Learning	20 Newsgroup database	Classified News	Technique which uses an unsupervised approach group unlabeled text document	Unavailability of specialized decision viz. unavailability of prediction of pipelining defect.
6	Shijian Lu . et.al.((2016)	Machine Learning	Input Document Image	Scene text recognition	System for scene text extraction.	Only few input parameters re used

7	Induja&Indu(2014)	Machine Learning	Text in Hindi, Nepali, Marathi, Bhojpuri or Sanskrit.	Output: Language identified for the given text.	Framework using frequent words and characters for language identification(LID)	Lack any explicit representation of long range dependency
8	Kavitha A.S (2015)	Machine Learning	Indus document	Text components pruning in the image	System for old historical text segmentation	Extending same methods for different Indian Scripts
9	Jan .R et al.(2014)	Machine Learning	Historical weather Documents	statistical analysis	System for reducing manual labeling	Time complexity is high
10	Morito & Sabourin (2019)	Machine Learning	Handwritten month word	Reduced Clusters on basis of Feature selection	Methodology for minimization of number of discriminant features	More input parameters should be included for clustering
11	Valy.D, (2016)	Machine Learning	Binary Images of Khmer palm leaf manuscript	Precision 99.528% Recall 99.534 % F-measure 99.531%.	System using Data clustering algorithm for text line extraction	Low extraction accuracy for handwritten cursive characters
12	Chen.K, (2015)	Machine Learning	Handwritten historical document images	Pixel accuracy, Mean pixel accuracy	System for text extraction of historical document using pixel labeling problem, instead of using preexisting features to train the classifier .	Less number of training images are used
13	Pei.W .et.al (2013)	Machine Learning		f -measure 71%	System for scene text detection was proposed using adaptive clustering	Lesser accuracy for multi-orientation cursive text.
14	Zhu.S & Zanibbi.R(2016)	Deep Learning	Scene Text dataset	Obtains pixel, character, and word detection f-measures of 93.14%, 90.26%, and 86.77%	Proposed an scene text detection system for natural scene	Lower accuracy with similar color in natural scenes.
15	R.S.Sabeenian.et.al, (2019)	Deep Learning	Tamil palm-leaf characters	Recognition Accuracy	Character recognition system using CNN	Accuracy is degraded by increasing the size of database.
16	Upadhaya&Jaiswal (2014)	Deep Learning	Sanskrit Sentences	Concatenate Segmented Text	Translation application using Artificial Intelligence	Translator does not work on tenses and web based thesis
17	Sharma&Ugwekar.A, (2018)	Deep Learning	Image containing printed Sanskrit characters	Articulation for a character in the form of English text .	Model based on image processing	More accuracy can be achived by using neural networks
18	T S Suganya & Dr. S Murugavalli	Machine Learning	Ancient Tamil script image	F Measure,Precision	Proposed a framework for feature extraction using shape and hough transform like Group Search Optimization and Firefly .	System incorporates only to ancient Tamil script, it can be extend for different language like devnagri, Hindi, Kannada, etc.
19	Ray.A. et.al, ((2015)	Deep Learning	Printed Oriya script	Reduced label error	Proposed a framework using RNN for Deep Bidirectional Long Short Term Memory(LSTM)	More training data is required to validate the model

20	In-Jung Kim and Xiaohui Xie	Deep Learning	Handwritten Hangul Characters	Rate of recognition on SERI95a and PE92	Recognizer for Handwritten Hangual recognition system based on Deep Convolutional Neural Networks (DCNN)	Achieve better accuracy on handwritten Cursive characters
21	Kesiman.M. et.al, (2016)	Deep Learning	Palm Leaf Manuscript Images	Character recognition Accuracy	Method for text and line segmentation without binarization process	To achieve better recognition rate of Balinese scrip
22	Maitra.D.et.al, (2015)	Deep Learning	Indian script characters or numerals	Recognition accuracies	System for recognition of various scripts handwritten characters.	Requires more better training on network architecture of specified dataset
23	Chen.K.et.al.(2017)	Deep Learning	Historical document images	Pixel accuracy, Mean pixel accuracy	Framework using Convolutional Neural Network (CNN) for page segmentation of historical document images using pixel	As number of layers increase performance accuracy decreases.
24	N.Sridevi&P.Subashini (2013)	Deep Learning	Hand engrossed Tamil Literature.	Feature vectors	Proposed a system for handwritten Tamil character recognition	Training and testing time can be further reduced.
25	Xiqun Zhang.et.al, (2017)	Deep Learning	Historical Tibetan documents	Precision, Recall, F-Measure	Framework for extracting text area from historical Tibetan text documents	Adapting historical Tibetan documents with different layout structures.
26	Huang .W .et.al, (2014)	Deep Learning	ICDAR 2011 benchmark dataset	F-measure reaches78.60%	System for Natural scene detection using CNN ,as Maximally Stable Extremal (MSERs) approach	Only foreground information (black ink) is relevant.
27	Katiyar.G and Mehfug.S, (2015)	Neural Network	CEDAR (Centre of Excellence for Document Analysis And Recognition).	Rate of recognition 93.23	System for recognizing feed forward neural network using hybrid feature extraction.	Achieve better accuracy on handwritten Cursive characters.
28	Wick.C & Puppe.F(2018)	Fully convolutional neural network (FCN)	GW5060, GW5064 and GW5066	Accuracy achived on Parzival data set 93.6% ,St. Gall 98.4%	Fully Convolutional neural network for historical document segmentation for processing single page using raw pixels rather than preprocessing which increases speed.	For further processing only black ink was relevant.
29	S.Bolan.et.a, (2013)	Image Processing	Handwritten-DIBCO 2010 degraded document image	F-Measure (%) PSNR NRM MPM	System for Image linearization, in order to carry out text segmentation containing external noise.	Still some more refinement need to be done on some DIBCO contests images.

Table-II: Study of various techniques along with the accuracy achieved

S.No	Technique	Author	Technique Used	Script type	Dataset	Accuracy
1	Supervised learning	A.Jeyapriya & C.S.Kanimozhi Selvi	Naive Based	Modern Script	ICDAR2013	80.36%
2		Bijalwan et al.(2014)	KNN	Modern Script	Reuters -21578	81.20%
3		Trstenjak.B (2014)		Modern Script	Milliyet_9c_1k, Hürriyet_6c_1k, 1150haber and Mini newsgroups	92.37%

4		Gopal&Raghav(2017)	SVM	Modern Script	Reuters 21578	97%
5		NiushaShafiabady (2015)		Modern Script	Reuters	95%
6		Shijian Lu . et.al.((2016)		Modern Script	ICDAR2013	F-measures 78.19 %
7	Unsupervised Learning	Induja&Indu (2014)	(N-gram)	Ancient Script Scrip	Word Net	90%
8		Kavitha A.S (2015)	K Means	Modern Script	Self-created dataset consisting of 500 images from archeology survey of India	93%
9		Jan .R et al.(2014)		Modern Script	MNIST database of handwritten digits	86.20%
10		Morito & Sabourin (2019)		Modern Script	Bangla numeral Devanagari numeral,Oriya numeral and Bangla basic character, Telgu Numerals ,English Numerals	86.60%
11		Valy.D, (2016)	Clustering	Ancient Script	Khmer Ancient Script palm leaf manuscripts	Precision 99.528% Recall 99.534 % F-measure 99.531%.
12		Chen.K (2015)	Layer by layer convolutional auto encoder.	Ancient Script	Saint Gall,he George	84.97%(10 K samples of Saint Gall)
13					Washington and Parzival	86.60%(10 K samples of he George Washington and Parzival)
14		Pei.W .et.al (2013)	Adaptive Clustering	Modern Script	MSRA-TD500	f -measure 71%
15		Zhu.S & Zanibbi.R (2016)	CNN Kmeans Clustering	Modern Script	CDAR 2015	Achieves pixel, character, and word detection f-measures of 93.14%, 90.26%, and 86.77%
16	Deep Learning	R. S. Sabeenian.et.al.(2019)	CNN	Ancient Script	Scanned dataset of palm leafs	96.1%.
17		Upadhaya&Jaiswal (2014)	Rule Based Machine Translation Using Artifical Intelligence	Ancient Script	Transhish (TTS)	94%
18		Sharma&Ugwekar.A(2018)	Radon and Euclidean distance transforms and ANN.	Ancient Script	Self-Generated dataset	94.117% On Constants and 76.47% On Vowels
19		Shi.B et al.(2017)	CRNN	Modern Script	Self-Generated data set	93%

20		Ray.A. et.al.((2015)	Recurrent Neural Network(RNN)	Modern Script	ICDAR 13	95.6%
21		In-Jung Kim and Xiaohui Xie	Deep Convolutional Neural Network (DCNN)	Ancient Script	SERI95 and PE92	95.96% on SERI95a and 92.92% on PE92
22		Kesiman.M. et.al.(2016)	Convolutional Neural Network (CNN)	Ancient Script		85%
23		Maitra.D.et.al.(2015)		Modern Script	Bangla numeral Devanagari numeral,Oriya numeral and Bangla basic character, Telgu Numerals ,English Numerals	99.1%
24		Chen.K.et.al.(2017)		Ancient Script	G. Washington, St. Gal, Parzival, CB55, CSG18, CSG863.	10K labeled training pixels, the accuracy is Saint Gall 84.97% and from 86.54% for George Washington and Parzival
25		N.Sridevi&P.Subashini (2013)	Extreme Learning Machine (ELM)	Ancient Script	Characters from Ancient Script Tamil Script	87%-95%
26		Xiqun Zhang.et.al.(2017)	Connected Characters (CC)	Ancient Script	The Complete Works of the Panchen Lama	F-measure reaches85.60%, Recall reaches 98.58%Precision evaluated is75.64%
27		Huang .W .et.al (2014)	Convolutional Neural Network (CNN)+MSER + sliding window	Modern Script	ICDAR2013	F-measure reaches78.60%
28		Wick.C & Puppe.F(2018)	Fully convolutional neural network (FCN)	Ancient Script	GW5060, GW5064 and GW5066	Parzival data set 93.6% ,St. Gall 98.4%
29	Image Processing	S.Bolan.et.al.(2013)	Canny Edge Detection	Modern Script	DIBCO-2010,DIBCO 2009 &2011	93%

The table II further examines the various parameters relevant to text extraction and comparison done on various techniques used within the domains like machine learning and deep learning along with dataset and accuracy achieved by each of them. The Table II further compares the accuracy level based on different type of dataset.

The strength of comparison lies in (1) understanding the challenges of research in text extraction (2) illustrating the effect on challenge accuracy (3) recognizing the usability of various text extraction techniques in different domains. Firstly, the comparison has been done on the first thirteen research papers. The target of the comparison is recognizing various techniques used for text mining using machine learning. It also delivers the effect through accuracy measures.

The second comparison has been done between the rest fourteen research papers. The target of the comparison is to recognize the accuracy level of extraction between modern and ancient script and as most of the modern scripts are available in the digitized and tagged format, therefore it is recommended that more study has to be done on ancient script.

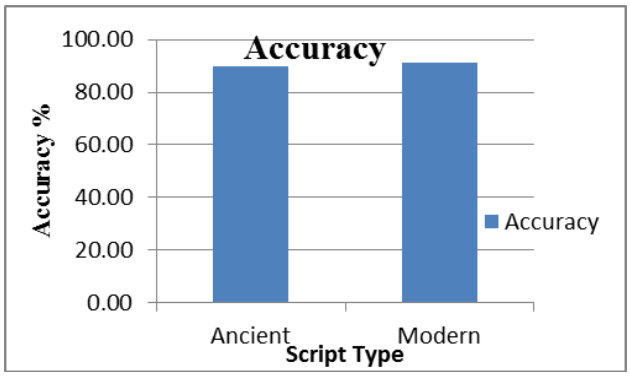

Fig.1 Accuracy (%)with various Script Type

Fig.1 presents the accuracy (%) comparison in modern and ancient script used in Table 2. Accuracy level achieved in ancient script is lower by 5% than modern script, as it contains noise and is not available in digitized and tagged format. Therefore, an extensive study has to be performed on ancient script [43].

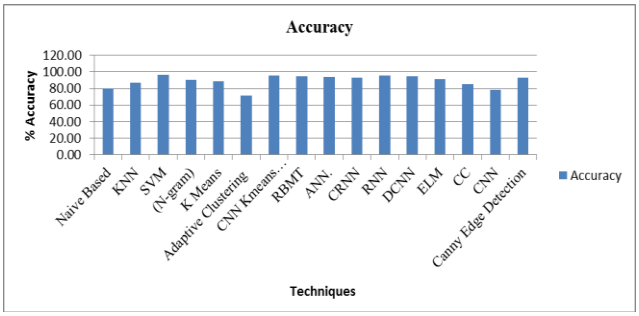

Fig.2 Technique Wise Accuracy (%)

Fig.2 presents the result accuracy in the form of (%) achieved with various machine and deep learning techniques for script text extraction in various studies as given in Table 2. A higher accuracy of approximately 3% has been achieved using hybrid model (CNN and KNN) in comparison to standalone techniques for script extraction.

III. CONCLUSIONS

The research concludes the effect of various text extraction methods using machine and deep learning techniques. The average accuracy (%) is being evaluated based on two comparisons done among twenty nine research papers. The first analysis has been done on the basis of language script used and the accuracy level has been found lower (5%) on ancient script, due to lack of digitized and tagged format. Moreover, the ancient script may contain lot of noise, which effects the average accuracy (%) achieved, on the basis of it can be concluded that further study may be done on ancient script extraction. The other study is been done on the basis of various machine and deep learning techniques used for text extraction. Therefore, it may be concluded that when standalone approach is used, it depreciates the average accuracy level for script extraction.

REFERENCES

1. A. A. Shah and K. Rana, "A Review on Supervised Machine Learning Text Categorization Approaches," 2018 Int. Conf. Circuits Syst. Digit.Enterp. Technol., pp. 1–6, 2019.
2. A. Ray, S. Rajeswar, and S. Chaudhury, "Text recognition using deep BLSTM networks," ICAPR 2015 - 2015 8th Int. Conf. Adv. Pattern Recognit., 2015.
3. À P. U., À U. C. J., and À K. A., "TranSish : Translator from Sanskrit toEnglish-A Rule based Machine Translation," vol. 4, no. 5, pp. 3463–3466, 2014.
4. A. S. Kavitha, P. Shivakumara, G. H. Kumar, and T. Lu, "Text segmentation in degraded historical document images," Egypt. Informatics J., vol. 17, no. 2, pp. 189–197, 2016.
5. B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," Procedia Eng., vol. 69, pp. 1356–1364, 2014.
6. B. Su and S. Lu, "Accurate recognition of words in scenes without character segmentation using recurrent neural network," Pattern Recognit., vol. 63, no. June 2016, pp. 397–405, 2017.
7. B. Su, S. Lu, and C. L. Tan, "Robust document image binarization technique for degraded document images," IEEE Trans. Image Process., vol. 22, no. 4, pp. 1408–1417, 2013.
8. C. L. Cameras, O. Enqvist, and F. Kahl, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 7, pp. 1455–1461, 2017.
9. C. Wick and F. Puppe, "Fully convolutional neural networks for page segmentation of historical document images," Proc. - 13th IAPR Int. Work. Doc. Anal. Syst. DAS 2018, pp. 287–292, 2018
10. D. Sen Maitra, U. Bhattacharya, and S. K. Parui, "CNN based common approach to handwritten character recognition of multiple scripts," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2015-Novem, pp. 1021–1025, 2015.
11. D . M. E. D. M. Hussein, "A survey on sentiment analysis challenges," J. King Saud Univ. - Eng. Sci., vol. 30, no. 4, pp. 330–338, 2018.
12. D. Valy, M. Verleysen, and K. Sok, "Line segmentation approach for ancient palm leaf manuscripts using competitive learning algorithm," Proc. Int. Conf. Front. Handwrit. Recognition, ICFHR, pp. 108–113, 2017.
13. F. Patel and N. Soni, "Text mining: A Brief survey," Int. J. Adv. Comput. Res., vol. 2, no. 4, pp. 243–248, 2012.
14. G. Katiyar and S. Mehfuz, "MLPNN based handwritten character recognition using combined feature extraction," Int. Conf. Comput. Commun. Autom. ICCCA 2015, pp. 1155–1159, 2015.
15. Joan Pastor-Pellicer, Muhammad Zeshan Afzal, Marcus Liwicki‡ & Mar´ıa Jose Castro-Bleda "Complete Text Line Extraction with Convolutional Neural Networks and Watershed Transform", p.p 30-35, 2016.
16. J. Richarz, S. Vajda, R. Grzeszick, and G. A. Fink, "Semi-supervised learning for character recognition in historical archive documents," Pattern Recognit., vol. 47, no. 3, pp. 1011–1020, 2014.
17. I. J. Kim and X. Xie, "Handwritten Hangul recognition using deep convolutional neural networks," Int. J. Doc. Anal. Recognit., vol. 18, no. 1, pp. 1–13, 2014.
18. K. U. Sharma and A. A. Ugwekar, "English transcription of sanskrit characters using predefined templates," Proc. 2017 2nd IEEE Int. Conf. Electr. Comput. Commun. Technol. ICECCT 2017, pp. 0–3, 2017.
19. K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, "Page segmentation of historical document images with convolutional autoencoders," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2015-Novem, pp. 1011–1015, 2015.
20. K. Indhuja, M. Indu, C. Sreejith, and P. C. R. Raj, "Text Based Language Identification System for Indian Languages Following Devanagiri Script," vol. 3, no. 4, pp. 327–331, 2014.
21. M. Morita, R. Sabourin, F. Bortolozzi, and C. Y. Suen, "Unsupervised feature selection using multi-objective genetic algorithms for handwritten word recognition," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2003-Janua, pp. 666–670, 2003.
22. M. W. A. Kesiman, J. C. Burie, and J. M. Ogier, "A new scheme for text line and character segmentation from gray scale images of palm leaf manuscript," Proc. Int. Conf. Front. Handwrit. Recognition, ICFHR, pp. 325–330, 2017.
23. M. Avadesh and N. Goyal, "Optical character recognition for sanskrit using convolution neural networks," Proc. - 13th IAPR Int. Work. Doc. Anal. Syst. DAS 2018, pp. 447–452, 2018.
24. M. Yousefi-Azar and L. Hamey, "Text summarization using unsupervised deep learning," Expert Syst. Appl., vol. 68, pp. 93–105, 2017
25. N. Kindo, G. Bhuyan, and R. Padhy, Computing and Network Sustainability, vol. 75. Springer Singapore, 2019.
26. N. Sridevi and P. Subashini, "Combining Zernike moments with regional features for classification of handwritten ancient tamil scripts using extreme learning machine," 2013 IEEE Int. Conf. Emerg. Trends Comput. Commun. Nanotechnology, ICE-CCN 2013, no. Iceccn, pp. 158–162, 2013.
27. N. Shafiabady, L. H. Lee, R. Rajkumar, V. P. Kallimani, N. A. Akram, and D. Isa, "Using unsupervised clustering approach to train the Support Vector Machine for text classification," Neurocomputing, vol. 211, pp. 4–10, 2016.
28. N. S. Panyam, V. L. Vijaya, R. K. Krishnan, and K. R. Koteswara, "Modeling of palm leaf character recognition system using transform based techniques," Pattern Recognit. Lett., vol. 84, pp. 29–34, 2016.
29. P. Mittal, V. Singhal, and P. S. S. Jain, "ICT Enabled Vehicular Safe Distance Modeling at Indian Unmanned Railway Level Crossings," SSRN Electron. J., 2019.
30. R. Kumar, P. Kumar, and V. Singhal, "A Survey: Review of Cloud IoT Security Techniques, Issues, and Challenges," SSRN Electron. J., 2019.
31. S. Daggumati and P. Z. Revesz, "Data mining ancient script image data using convolutional neural networks," ACM Int. Conf. Proceeding Ser., pp. 267–272, 2018.
32. S. Gopal and S. Raghav, "Automatic document retrieval using SVM machine learning," Proc. 2017 Int. Conf. Smart Technol. Smart Nation, SmartTechCon 2017, pp. 896–901, 2018.

33. S. Lu, T. Chen, S. Tian, J. H. Lim, and C. L. Tan, "Scene text extraction based on edges and support vector regression," Int. J. Doc. Anal. Recognit., vol. 18, no. 2, pp. 125–135, 2015.
34. S. Mirza, "Design and Implementation of Predictive Model for Prognosis of Diabetes Using Data Mining Techniques," Int. J. Adv. Res. Comput. Sci., vol. 9, no. 2, pp. 393–398, 2018.
35. S. Sudholt and G. A. Fink, "PHOCNet: A deep convolutional neural network for word spotting in handwritten documents," Proc. Int. Conf. Front. Handwrit. Recognition, ICFHR, pp. 277–282, 2017.
36. S. Zhu and R. Zanibbi, "A text detection system for natural scenes with convolutional feature learning and cascaded classification," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 625–632, 2016.
37. V. Singhal, S. S. Jain, and M. Parida, "Train sound level detection system at unmanned railway level crossings," Eur. Transp. - Trasp. Eur., no. 68, 2018.
38. V. Singhal and S. Jain, "Safety Information System of Indian Unmanned Railway Level Crossings," IOSR J. Mech. Civ. Eng. Ver. III, vol. 12, no. 4, pp. 2278–1684, 2015.
39. V. Bijalwan, V. Kumar, P. Kumari, and J. Pascual, "KNN based machine learning approach for text and document mining," Int. J. Database Theory Appl., vol. 7, no. 1, pp. 61–70, 2014.
40. V. Singhal and S. S. Jain, "Road driver behaviour evaluation at unmanned railway level crossings," Eur. Transp. - Trasp. Eur., 2017.
41. W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced MSER trees," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8692 LNCS, no. PART 4, pp. 497–511, 2014.
42. X. C. Yin, W. Y. Pei, J. Zhang, and H. W. Hao, "Multi-Orientation Scene Text Detection with Adaptive Clustering," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1930–1937, 2015.
43. X. Zhang, L. Duan, L. Ma, and J. Wu, "Text extraction for historical tibetan document images based on connected component analysis and corner point detection," Commun. Comput. Inf. Sci., vol. 772, pp. 545–555, 2017.
44. Z. Lei, S. Zhao, H. Song, and J. Shen, "Scene text recognition using residual convolutional recurrent neural network," Mach. Vis. Appl., vol. 29, no. 5, pp. 861–871, 2018.

AUTHORS PROFILE

Shikha Verma has received her B. Tech.(CSE) from ,GLAITM, Mathura, M.Tech.(CSE) from JMIT, Radur and pursuing Ph.D. from Jaipur National University, Jaipur. She has overall teaching experience of more than 16 years and currently working as Assistant Professor at JSS Academy of Technical Education, Noida since 2006. Her area of interest is Machine learning, Data mining, Operating system .

Dr. Sonu Mittal is working as an Associate Professor and Coordinator Ph.D. Program in Computer Science & Engineering Department of Jaipur National University, Jaipur. He is UGC-NET qualified and has more than 16 years of teaching and research experience. He has guided 05 Ph.D. research works. Dr. Mittal has authored a book on computer networks and published 28 papers in international journals and conferences. He is member of various professional bodies like ACM and CSI. He has chaired 5 international/ national conferences and has been actively engaged as reviewer for various national and international forums. His major areas of research interest are Software Engineering, Computer Networks and Machine Learning.

Dr. Vivek Singhal is working as Associate Professor at JSS Academy of Technical Education, Noida. He has done his Ph.D. from IIT Roorkee on the topic "ICT Based Road Vehicle-Train Collision Avoidance System at Unmanned Railway Level Crossing". Roorkee. He has vast teaching and research experience of around 11 years. He has received gold medal from Vaishya Samaj, Meerut for excellent performance in M.Tech. He has been a member of several academic and administrative bodies like CSI, IRC etc. He has published many research papers in different referred journals and conferences. His area of research includes Wireless Networks, Congestion Control, Machine Learning, Intelligent Transportation systems, Rail-Road Safety and Geographical Information Systems.