ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

Different Feature and Classifiers the Study of Automatic Handwritten Devnagari Text Generation

Mr.Vajid Khan[1], Dr.Yogesh Kumar Sharma[2]

Research Scholar, Department of CSE, Shri JJTU, Rajasthan, India [1]

Professor, Head of Research Coordinator, Department of CSE, Shri JJTU Rajasthan, India[2]

ABSTRACT*:* In recent years analysis towards Indian written character recognition is obtaining increasing attention. Many approaches are planned by the researchers towards handwritten Indian character recognition and lots of recognition systems for isolated written numerals/characters area unit on the market within the literature. To get idea of the popularity results of various classifiers and to provide new benchmark for future analysis, during this paper a comparative study of Devnagari written character recognition victimization twelve completely different classifiers and 4 sets of feature is bestowed. Projection distance, mathematical space technique, linear discriminant operate, support vector machines, modified quadratic discriminant operate, likeness learning, geometrician distance, nearest neighbour, k-Nearest neighbour, changed projection distance, compound projection distance, and compound changed quadratic discriminant operate area unit used as completely different classifiers. Feature sets utilized in the classifiers area unit computed supported curvature and gradient info obtained from binary as well as gray-scale pictures.

KEYWORDS: Handwriting recognition, CNN-RNN network, Data augmentation, Image pre-processing.

I. INTRODUCTION

In the gift digital world, it's become obligatory to own all the out there info in a very digital type recognized by machines. In country like Asian nation, wherever there's abundance of knowledge within the style of manuscripts, ancient texts, books, etc. that are historically out there in written and written type, such written material are in-adequate once it involves looking info among thousands of pages. It's to be digitized and regenerate to a matter type in-order to acknowledge by machines doing searches of 1,000,000 pages/second. We tend to even have to retrieve and extract the text that is deteriorated. Then only, actuality information of Indian history, tradition and culture would be out there to the plenty and also the digital revolution would be aforementioned to own reached the data age.

In the field of pattern recognition and computing Optical Character Recognition (OCR) has become one among the foremost undefeated applications. Today, moderately economical and cheap OCR packages are commercially out there to acknowledge written texts in wide used languages like English, Chinese, and Japanese. OCR is that the most essential a part of Document Analysis System that converts the scanned pictures of text, books, magazines, and newspapers into electronic text. The method of Document Analysis Recognition is divided into 2 components, namely, written and handwriting character recognition. The written documents will more be divided into 2 parts: smart quality written documents and degraded written documents. Degradation of the text ends up in touching, broken, significant written, typed, faxed document, back facet visible, manually underlined lines of text and digital image from camera figure 1.1

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

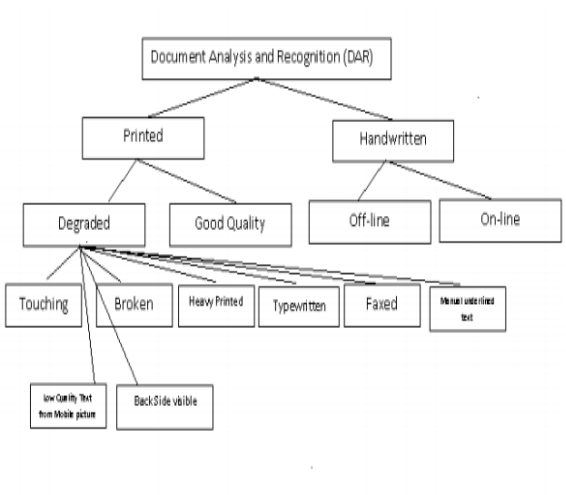

Figure: 1. Categories of DAR

Recognition of degraded documents has been an energetic analysis space within the field of pattern recognition. Handwriting character recognition has been divided into offline and on-line character recognition, as shown in figure 1.1. Offline documents area unit scanned pictures of prewritten text, usually on a sheet of paper. In on-line handwriting recognition, information area unit captured throughout the writing method with the assistance of a special pen And an electronic surface. All Indian scripts area unit cursive in nature that makes them laborious to acknowledge by machines. Scripts like Devnagari script, Gujarati, Bengali and plenty of others have conjuncts or joint-characters increasing segmentation difficulties. to feature to it, varied fonts of assorted sizes used for printing texts over the years, the standard of paper, scanning resolution, pictures in texts and degraded quality of text and paper etc. puts challenges for the researchers for image process job. Also, it needs immense linguistic power to use post-processing.

Recognition of degraded documents has been an energetic analysis space within the field of pattern recognition. Handwriting character recognition has been divided into offline and on-line character recognition, as shown in figure one.1. Offline documents area unit scanned pictures of prewritten text, usually on a sheet of paper. In on-line handwriting recognition, information area unit captured throughout the writing method with the assistance of a special pen and an electronic surface. All Indian scripts area unit cursive in nature that makes them laborious to acknowledge by machines. Scripts like Devnagari script, Gujarati, Bengali and plenty of others have conjuncts or joint-characters increasing segmentation difficulties. To feature to it, varied fonts of assorted sizes used for printing texts over the years, the standard of paper, scanning resolution, and pictures in texts and degraded quality of text and paper etc. puts challenges for the researchers for image process job. Also, it needs immense linguistic power to use post-processing.

Devanagari is that the script used for writing several official languages in Bharat like Hindi, Marathi, Sindhi, Nepali, Sanskritic language and Konkani, wherever Hindi is that the national language of the country. Hindi is additionally the third most well liked language within the world. Many alternative Indian languages like Gujarati, Punjabi, and Bengali use scripts almost like Devnagari script. Over five hundred million individuals use syllabary for documentation in central and northern components of Bharat.

Recognition of handwritten characters has been a popular research area for many years because of its various application

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

potentials. Some of its potential application areas are Postal Automation, Bank cheque processing, automatic data entry, etc. In recent years research towards Indian handwriting character recognition is getting increasing attention. Many approaches have been proposed by the researchers towards handwritten Indian character recognition and many recognition systems for isolated handwritten numerals/characters are available in the literature.

To get idea of the recognition results of different classifiers and to provide new benchmark for future research, in this paper a comparative study of Devnagari handwritten character recognition results is reported here. To compare the performance, twelve different classifiers and four different features computed from gradient and curvature information of the binary as well as gray-scale images are used here. Classifier like Projection distance (PD), Subspace method (SM), Linear discriminant function (LDF), Support vector machines (SVM), Modified quadratic discriminant function (MQDF), Mirror image learning (MIL), Euclidean distance (ED), Nearest neighbour, k-Nearest neighbour (k-NN), Modified Projection distance (MPD), Compound projection distance (CPD), and Compound modified quadratic discriminant function (CMQDF) are considered.

II. LITURATURE SURVEY

Devnagari is the most popular script in India and the Indian national language, Hindi, is written in Devnagari script. Nepali, Sanskrit and Marathi are also written in Devnagari script. Moreover, Hindi is the third most popular language in the world .The alphabet of the modern Devnagari script consists of 14 vowels and 33 consonants. These characters are called basic characters. The shape of basic characters of Devnagari script is shown in Fig.1 and we used these 47 basic characters for our experiment. Writing mode in Devnagari script is from left to right. The concept of upper/lower case is absent in Devnagari script. In Devnagari script a vowel following a consonant takes a modified shape. Depending on the vowel, its modified shape is placed at the left, right (or both) or bottom of the consonant. These modified shapes are called modified characters. A consonant or vowel following a consonant sometimes takes a compound orthographic shape, which we call as compound character. Compound characters can be combinations of two consonants as well as a consonant and a vowel. Compounding of three or four characters also exists in the script. There are about 280 compounds
Characters in Devnagari

Figure 2: Samples of handwritten Devnagari basic characters (a) Vowels (b) Consonants. To get an idea about the shape of the character, samples of printed characters are shown in the left side of the respective handwritten characters.

Devanagari script started in early 1970s. An extensive research on printed Devanagari text was carried out by Veena Bansal and R. M. K. Sinha. First system for hand-written numeral recognition of Devanagari characters was proposed by

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

R. Bajaj et al. P. M. Patil and T.R. Sontakke also presented an algorithm for handwritten Devanagari numeral recognition which was rotation, scale and translation invariant. U. Pal et al presented a system for off-line handwritten character recognition of Devanagari using directional information for extracting features. A technique for accuracy improvement of Devanagari character recognition system was proposed by U. Pal et al using two features based upon directional and curvature information in the characters and applied to the combination of two classifiers namely, support vector machines and modified quadratic discriminate function. A comparative study of various features and classifiers used for handwritten Devanagari character recognition was done by U. Pal et al. N Sharma et al. proposed a method where the features are obtained from the directional chain codes information of the contour points.
A Multi-feature multi-classifier scheme for handwritten Devanagari characters is proposed by S. Shelke and S. Apte, which combines neural network and template matching recognition Approaches. Recognition of Non-Compound Handwritten Devnagari characters using a Combination of MLP and Minimum Edit Distance was proposed by S. Arora et al. for compound characters involving two stages. A methodology for off-line isolated handwritten Devanagari character recognition is proposed by Mahesh Jangid. Brijmohan Singh and Ankush Mittal proposed the two different methods for extracting features from handwritten Devnagari characters, the Curvelet Transform and the Character Geometry

III. PROPOSED METHODOLOGY

Handwritten recognition and text generation is currently getting the attention of researchers because of possible applications in assisting technology for blind and visually impaired users, human–robot interaction, automatic data entry for business documents, etc. Here, I have intended to develop Ant Miner Algorithm (AMA) for recognition and text generation of the Devanagari scripts. The proposed Devanagari text generation system using AMA is designed to accept scanned document images which pre-processes, segments, feature extraction and lastly recognition and text generation. Generally the AMA architecture is consists of two phases such as training and testing phases. During the training phase, a known set of text document images are taken which are further processed to get Devanagari characters. When top lines are cleared from words during segmentation process, then independent and isolated Devanagari characters are obtained. In the segmentation process, the Devanagari characters can either be non-modified simple characters or those characters whose modifiers are separated. After that, the segmented character features are extracted for training the AMA algorithm. Initially, the AMA is calculated the minimum distance among the character and its upper/right/left modifier. After that, the recognition and text generation is attained with the help of AMA algorithm. During testing phase, the performance of the AMA algorithm is tested with an unknown set of document images. The proposed method will be compared with the existing methods of Artificial Neural Network (ANN) and Deep Learning (DL).

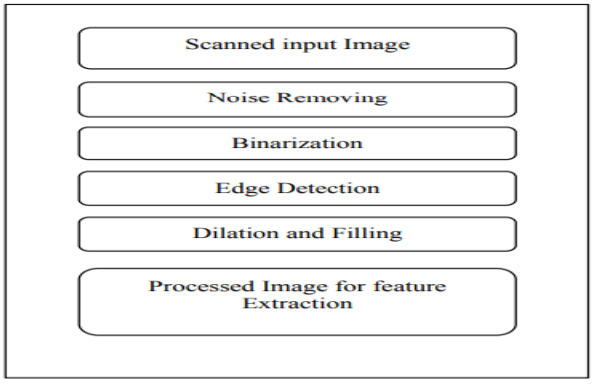

Fig 2: Stage of Pre-Processing

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

Process of the proposed methodology

Step 1: Input image of Devanagari Script handwritten and scanned images are collected from the open source system.
Step 2: Pre-processing Stage: Noise removal, Banalization and Skew correction
Step 3: Segmentation: Line segmentation, word segmentation and character segmentation
Step 4: Feature extraction: Linear discriminant Analysis (LDA), Scale Invariant Feature Transform (SIFT), Local Binary Patterns (LBP) and Discrete Cosine Transform (DCT).
Step 5: AMA algorithm is working based on the two phases such as training phase and testing phase. In training phase, the features are training after that testing phase which can be tested.
Step 6: Finally, Devanagari handwritten recognized and text generation is achieved with the help of the AMA algorithm.
Simulation Tool: Python

IV. CONCLUSION

To get the concept of the popularity results of various classifiers and to supply new benchmark for future analysis, during this paper a comparative study of Devnagari written character recognition exploitation twelve totally different classifiers is reported here. Results of various classifiers are mentioned and justifications of the results obtained are briefed. We have a tendency to note that likeness Learning gave overall higher results among the classifiers and shown highest results accuracy on grey-scale curvature options.

REFERENCESS

1. Shailendra Kumar Shrivastava,Sanjay Gharde," Support Vector Machine for Handwritten Devnagari Numeral Recognition",International Journal of Computer Applications,Volume 7,No.11,October 2010.
2. T. M. Rath and R. Manmatha, "Word spotting for historical documents," IJDAR, 2007.
3. S. N. Srihari and E. J. Kuebert, "Integration of handwritten address interpretation technology into the united states postal service remote computer reader system," in DAS, 1997.
4. R. J. Milewski, V. Govindaraju, and A. Bhardwaj, "Automatic recognition of handwritten medical forms for search engines," IJDAR, 2009.
5. G. F. Simons and C. D. Fennig, "Ethnologue: Languages of the world," SIL International, 2017.
6. U.-V. Marti and H. Bunke, "The iam-database: an english sentence database for offline handwriting recognition," IJDAR, 2002.
7. U. Pal and B. Chaudhuri, "Indian script character recognition: a survey," PR, 2004.
8. U. Stiehl, "Sanskrit-kompendium," Heidelberg: Hüthing, 2002.
9. U. Pal and B. Chaudhuri, "Indian script character recognition: a survey," PR, 2004.
10. S. Arora, D. Bhattacharjee, M. Nasipuri, L. Malik, M. Kundu, and D. K. Basu, "Performance comparison of SVM and ANN for handwritten devnagari character recognition," arXiv preprint arXiv:1006.5902, 2010.
11. K. Mehrotra, S. Jetley, A. Deshmukh, and S. Belhe, "Unconstrained handwritten devanagari character recognition using convolutional neural networks," in 4th International Workshop on Multilingual OCR, 2013.
12. P. P. Roy, A. K. Bhunia, A. Das, P. Dey, and U. Pal, "HMM-based indic handwritten word recognition using zone segmentation," PR, 2016.
13. B. Shaw, U. Bhattacharya, and S. K. Parui, "Combination of features for efficient recognition of offline handwritten devanagari words," in ICFHR, 2014.
14. B. Shaw, S. K. Parui, and M. Shridhar, "Offline handwritten devanagari word recognition: A holistic approach based on directional chain code feature and HMM," in ICIT, 2008.
15. A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," PAMI, 2009.
16. N. Sankaran, A. Neelappa, and C. V. Jawahar, "Devanagari text recognition: A transcription based formulation," in ICDAR, 2013.
17. R. Jayadevan, S. R. Kolhe, P. M. Patil, and U. Pal, "Database development and recognition of handwritten devanagari legal amount words," in ICDAR, 2011.
18. A. Alaei, U. Pal, and P. Nagabhushan, "Dataset and ground truth for handwritten text in four different scripts," IJPRAI, 2012.
19. R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, "Cmaterdb1: a database of unconstrained handwritten bangla and bangla–english mixed script document image," IJDAR, 2012.
20. S. Thadchanamoorthy, N. Kodikara, H. Premaretne, U. Pal, and F. Kimura, "Tamil handwritten city name database development and recognition for postal automation," in ICDAR, 2013.
21. B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

recognition," PAMI, 2016.
22. M. Jaderberg, K. Simonyan, A. Zisserman et al., "Spatial transformer networks," in NIPS, 2015.
23. A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in ICML, 2006.
24. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in CVPR, 2015
25. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
26. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016.
27. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in ICML, 2015.
28. V. Pham, T. Bluche, C. Kermorvant, and J. Louradour, "Dropout improves recurrent neural networks for handwriting recognition," in ICFHR, 2014.
29. W. Liu, C. Chen, K.-Y. K. Wong, Z. Su, and J. Han, "Star-net: A spatial attention residue network for scene text recognition." in BMVC, 2016.
30. P. Krishnan and C. V. Jawahar, "Generating synthetic data for text recognition," arXiv preprint arXiv:1608.04224, 2016.
31. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in ECCV, 2014.
32. T. Bluche and R. Messina, "Gated convolutional recurrent neural networks for multilingual handwriting recognition," in ICDAR, 2017.