Devnagari Script Categorization by Utilizing CNN and KNN

Sarika T. Deokate, Nilesh J. Uke

Abstract*: In the present days of digitized era, the work is carried out on the document analysis, categorization and dealing out it digitally. Therefore, in this work we introduced the model, which works well with the categorization of the Devnagari script-Marathi. We proposed this investigation for the Type-scripted manuscripts and tested on some handwritten manuscript. As spurious images may not generate the superior ending product, so it necessitated eradicating the noise and so utilized the Gaussian Approach with Otsu's approach. Once pre-processing is accomplished, the fragmentation of the content of the manuscript image executed. To excerpt the lines, the morphological manoeuvres utilized in this work and generated the superior consequences for the manuscripts. Similarly, the individual words are taken out from the line images, by utilizing the horizontal projection approach. However, to excerpt the character/symbols from the words, we utilized the combination of shirorekha removal with vertical projection scheme. The tolerance factor estimated to fragment the characters perfectly. KNN and CNN categorizers utilized to categorize the characters. The result illustrated for the varying size of the dataset. We got average 96% of precision for the KNN categorizer, when evaluated on the test and trained dataset with k=1 to 5. For CNN categorizer the outcome we got is with 100% of precision for the trained and validated dataset. Then we processed the letters with proper post processing by utilizing the Unicode approach. The transformation simplified through the construction and respective labeling of our created dataset. This study can be utilized further for the visor and disabled people. In further study, concentration will be on the generalized approach for both type-printed and handwritten manuscripts.

***Index Terms*: CNN, Image Enhancement, Document Analysis, KNN .**

I. INTRODUCTION

The field of image processing is expanding swiftly and day by day striving to solve the complex problems proficiently with ease of working. By stages, researchers are exploring for the finer solution and vamping the solution for each business. Varied approaches are designed for Thresholding, detecting edges, fragmentation, pattern identification, classification for all sorts of applications. With conventional approaches, the nature-inspired algorithms are also utilized to obtain the optimal way out. The nature's theory, concept, way of working is adapted in digital era to find the optimal solutions to huge set of complex problems[1]–[3].

Revised Manuscript Received on March 10, 2019.

Sarika T. Deokate, Research Scholar, Department of Computer Engineering, JJTU Rajasthan, India.

Dr. Nilesh J. Uke, Principal, Department of Computer Engineering, TAE Pune, Pune, India.

One of the disciplines in digital image processing taken into consideration is OCR i.e. optical character recognition. Conventional manuscripts are strenuous to preserve for far time and even is degraded due to several environmental issue[10]. The manuscript can be identified in two ways, one in which individual character are identified one by one, second in which words are identified wholly.

The Russian scientist Tyuring initiated the OCR as thought in early 1900. He developed his first design for the visor people. However, it was quite slow and costlier to carry out the task. The initial work of automatic character identification was concentrated mostly on these machine printed manuscript and extremely limited to Latin script, which utilized the statistical identifiers and template matching for the hand-written and printed manuscript respectively. IBM initiated the optical reader, IBM 1418 and IBM 1428 with their ever first design, which had capability to interpret the printed and hand-printed figures/numbers. Postal mechanization performed by US Postal services to construe the postal codes to acquire the addresses of the destination in mid stage of 1960. In the era of 1990's the Image Processing utilized with the approach of artificial intelligence and recognition of Patterns. The further powerful and proficient approaches has been developed by utilizing the theory of HMM (Hidden Markov model), NN (Neural Network), Fuzzy reasoning and other such techniques for both on-line and off-line applications[4]–[6]. Devnagari script identification for the printed manuscript, initiated in 1970's era. Syntactic pattern investigation was utilized for the well-established image script for Devnagari machine printed manuscript [7].

II. LITERATURE SURVEY

Before supplying the extracted manuscript letters images for the feature extraction, it is vital to intensify the image. Manuscript image gets debased at the time of scanning or some of the manuscripts are very old. So before initiating towards the feature extraction and post-processing, these images must be smoothen with eradication of noise, skew discovery and modification converting to bi-level for betterment in outcome[8], [9].

Nowadays tremendous research is carried on document analysis and categorization digitally. Old manuscripts like magazines, novels, study material, business cards, and post cards can be straightforwardly construed to computer via these OCRs. Fragmentation and Identification of the handwritten or typed lettering is thorny due to a variety of aspect. Words need to be fragmented perfectly and with more precision,

else the recognition rate decreases. In this application, segmentation can be done at line, words and character level for variety of scripts. If characters are partially fragmented, still the outcome of the categorization degrades. Even due spurious entities, broken noise amid the manuscript may degrade the quality of the manuscript. Lot of work carried out for the superior fragmentation[13].

Text line or header line identification and elimination are essential to shun the mystification amongst the original text lines and background line. Header line or Shirorekha can be identified and disconnected to separate out upper zone and lower zones. These lines also eliminated to segment the words and characters precisely. In the past, lot efforts are done to discover and eliminate the lines[10]. Normally used techniques are Hough transform, histogram profile projection[11]. Gap between words is large as compared to characters. The extracted line structure is used as input here. For word segmentation vertical histogram profile, boundary box, windowing, regionprops functions are the well-known approach used by many researchers[12]–[16].

As we obtain the detached words from the manuscript it is vital task to break up the characters for identification and then individually categorized. It is required to provide the schemes, which remove, overlie or coupled characters distinguishing the manuscript text and graphical contents, by eradicating noise from manuscript text. Feature extraction is utilized to categorize the features, so that appropriate approach can be applied to extract the image features/objects. The achieved features are then used in classification procedure as input. In feature extraction, selection of the right feature is also important task.In feature extraction, the global or local neighborhood technique is utilized to discover the feature region e.g. color, shape, statistical properties, size, solidity, Euler number, filled region etc. [17]. Manuscript identified characters need to be categorized to interpret it truthfully. A variety of approaches have been utilized and executed proficiently for the categorization of the fragmented letters. These approaches performed the superior results with some varying conditions. A novel approach has been designed for discovering and identifying the transcript entities from the compound images and frames of the videos. Text discovery is performed the segregated transcript lines. The images containing the text entities are fragmented multiple times.

U. Pal, Wakabayashi, Sharma, and Kimura has designed a model, which utilized to identify the six distinct handwritten numerals of Indian scripts [18]. S. Arora et.al. has used four feature excerption techniques for hand written Devnagari character identification i.e. shadow feature, intersection, straight line fitting technique and chain code histogram with outcome of 92.80% [19]. Various researchers designed an approach for identification of Devnagari script. To test the performance NN, reduced NN, KNN Euclidian distance based KNN and other similar variant classifiers are utilized. Unconstrained manuscript handwritten Marathi characters was utilized by [15].

To solve the composite problems, the machine learning conceptions are adapted rapidly. Deep learning is the burning idea utilized in the computer vision. Author presented the detailed architecture of the deep learning. Deep learning does not need the individual extraction of feature as normal machine learning systems. It performs the extraction of the image features as the one of the built-in functionality of the categorization. CNN model is intended in such a ways that it acclimatizes the multilayer perception and necessitated very limited pre-processing [20],[23].

III. PROPOSED METHODOLOGY

Input: Manuscript Images
Output: Pre-processed Manuscript image

Algorithm:

- Manuscript Image Scanning
  - Photograph / Scanned Image
- Binarization of Scanned Manuscript Image
  - Color image to Bi-level transformation
- Noise Eradication
  - Noise discovery and eradication by utilizing the morphological filters.
  - Removed tiny Image Entities
- Skew Discovery and Correction
- Resizing
- Normalization of the fragmented images, trained and testing dataset
- Fragmentation
  - Line level Fragmentation
  - Word level Fragmentation
  - Character level Fragmentation

To perform the fragmentation of lines, characters and words, the existing histogram approach with morphological operations has been utilized, but with some enhanced approaches.

In this fragmentation, instead of fragmenting the characters, upper modifiers and lower modifiers separately, the characters are separated with the modifiers. To check the perfect characters even Shirorekha i.e. header line is eradicated and bounding region is estimated.

Morphological operations are applied in computer vision for the several applications like smoothing, thinning, shape identification, noise eradication, contour and edge discovery, object discovery and analysis. These operations are utilized for detecting the objects, lines, column discovery. In this study we utilized the erosion with dilation operation and horizontal profile for detecting the manuscript lines.

As we have fragmented the whole characters with modifiers, we prepared our dataset for the printed one by using the typed letter with modifiers and done the completely process and preprocessed with our system for further utilization in categorization. We also collected the samples from the newspapers and other resources to enhance the dataset.

- Identification and Categorization
  - Dataset Preparation
  - Labeling of the dataset
  - KNN approach
  - Deep Learning

Classifier is in essence a system utilized for mapping of input space to the given class space. Diverse classifier presents the diverse outcome for the classification. In our system, we analyzed and utilized the two categorizers one is KNN and second is CNN-Deep learning approach. This system has been illustrated in the following Figure.

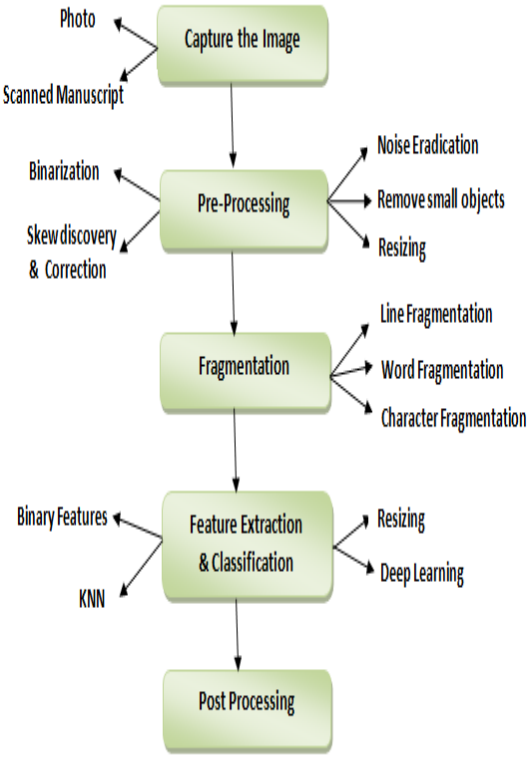


Fig. 1 Document Analysis

IV. RESULT AND DISCUSSION

The performance of every approach has been evaluated with considering the necessary parameters at each stage of the work. At the very first, we performed the augmentation of the source image by eradicating the spurious contents, tiny objects and unwanted objects from this source image. We assessed it on graphitic images, newspaper articles with some pictures in it. The precision of extracting the lines is depending on the quality of the images. In this system, if images are having the object in it, it is not considered at the time of fragmentation and eradicated it by considering it as unwanted part. We utilized as much possible type of images and got the outcome of 95% on an average. The remaining lines are fragmented partially, means 2 or 3 lines gets fragmented together.

However, once we apply the word fragmentation on these partially fragmented lines it takes out all the lines successfully with precision of 100%. So we can say that line fragmentation is done with 100% precision in our proposed system. It also works well for both printed and handwritten manuscripts with tiny inclination. If inclination is high, it is needed to correct the inclination. In similar way, the word fragmentation is providing the precision on an average of 97%. Again, our enhanced fragmentation approach fragments these fused words at character level fragmentation and takes out the characters. So again here for word fragmentation, the precision boosted from 97% to approximately 99 to 100%.

For our character fragmentation, we are getting approximately 85% characters fragmented correctly. Some of the character may not be fragmented correctly and may remain partially together. This is happening due to the crossing region of one character in the other character or may be connected to each other or some noisy interruption. Still we are working on the improvement of the same.

We prepared our dataset for the assessment of the categorization approaches and then evaluated the categorization effectively with greater precision. Still working on it, to increase and making perfect for the possibly all the fonts. We done the labeling in such way that the overhead of transformation of these recognized letter to soft form become an easy task. We utilized the standard Unicode codec system to encode and decode these letters at the time of post-processing.

We assessed the KNN extensively over the number of images for varying value of k. Initially we tested the outcome for the tiny quantity images and we got the 100% accuracy for k=1 but as we increased the value of k, the precision of the outcome is decreased. So gradually, we increased the dataset to suit for the categorization purpose. We utilized the distinct fonts to preserve it in the dataset to provide the more accuracy. We enhanced the dataset gradually to accomplish the requirement of the system.

In the following table 1 and figure 2, we illustrated the performance of the KNN for the varying size of the k for the training dataset and test dataset. As we can observe, the precision is diminishing with increasing value of k. For some of the letters, it is giving précised outcome for the value 1 or 2 or 3 or varying size depending on the availability of the pattern in the dataset.

Table I KNN outcome for the dataset of 15600 images

Value of K	*Accuracy*
1	99.94
2	97.41
3	96.53
4	93.37
5	92.92
6	91.58
7	91.32
8	90.58
9	90.47

The same table depicted in terms of the graph illustration.

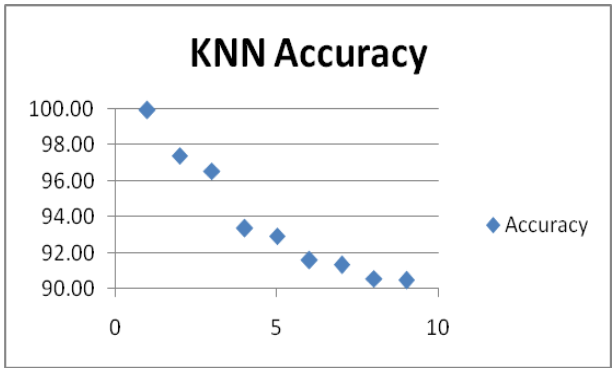
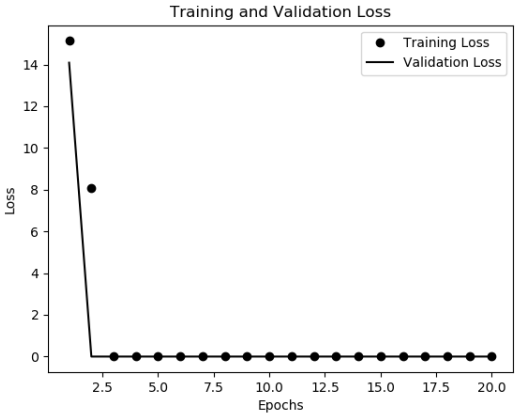


Fig. 2 KKN Precision Graph

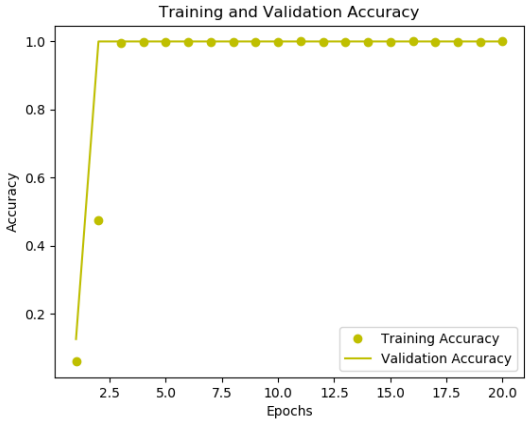
But when we actually evaluated this approach for the fragmented characters it is giving the outcome based on the quantity of the images utilized in the dataset. We assessed the KNN for the distinct letter, which resemble with each other may get confused to identify it correctly. We thoroughly observed the outcome for the क and फ़ व and ब, which resemble with each other a lot. We evaluated it with the varying size of the dataset. For this we utilized the distinct font style like Shivaji 01, 02, 05, Kruti Dev 55, Kruti Dev021 and similar other fonts to check the assessment. However, once we boosted the dataset, the precision is improved to 100%. For the increasing value of the k, some of the letter may be misidentified for some letter. As we showed in figure 2, we got average 96% of accuracy for the trained dataset for the varying K.

Next approach utilized is CNN utilizing the Keras[21] and Tensorflow[22]. The design was implemented in python with Conda settings. We assessed the CNN model with three convolutional levels followed by Max pooling and Drop-out. Two dense levels utilized. We checked the performance from the small size dataset to the varying size bit large dataset. We first evaluated our approach on 99 dataset and we got training accuracy of 100%. Out of it 89 was utilized for the training set and 10 for the validate set with epoch of quantity 20. Then we gradually augmented our dataset to 2440 with 2196 as training and 244 validate set.

With the same levels, we got 100% accuracy and very tiny over-fitting which is very negligible. As our problem is a multiclass categorization, it is suitable to utilize the RmsProp optimizer with the categorical cross_entropy to ground the model perfectly. RmsProp tailored to new learning proportion intended for every utilized parameter.



a) Loss for Drop=0.4



b) Accuracy for Drop=0.4

Fig. 3 Loss and Accuracy of trained and validated set

Next we altered the drop from 0.5 to 0.4 evaluated the system for the performance and the result is depicted in the figure 3.a and 3.b. For this set of factors, we got the accuracy of 100% for training and validate precision. But as compared to the parameters which we finally utilized in our model these factors are low performing with vary bit variance at the loss and accuracy. Still we are working on the dataset augmentation for the perfection in the outcome.

Further, samples utilized for the training and validation are split and performed 20 epochs for the entire dataset of quantity 15850. Trained on 40365 samples, validate on 4485 samples. The loss occurred at every epoch is diminished and precision is boosted. The model tallies this behavior at every epoch by examining the parameters and the outcome at the moment.

Therefore, if getting the lot much loss at some epoch does not mean that same behavior carried forward to the subsequent epochs. We can get the excellent outcome over the next processing epoch as compared to the previous level. In this work, we thus utilized the distinct parameters and utilized the optimal parameter set for our system.

V. CONCLUSION

The categorization and final processing of the manuscript performed in this work. The outcome of the proposed model illustrated here for the KNN and Deep Learning Conception CNN. The character fragmentation is tedious task in any document analysis system and we outperformed by fragmenting the letter with its modifiers. For performing this we have built our dataset. The outcome has been assessed and evaluated in the last section. We are working towards getting, the perfect precision at every level and will try to create a generalized model for the Printed as well as handwritten manuscripts in future.

REFERENCES

1. M. Tuba, "Multilevel image thresholding by nature-inspired algorithms : A short review ," *Icisp*, vol. 22, no. 3, pp. 318–338, 2014.
2. Yongsheng Pan, Tao Zhou, and Yong Xia, "Bacterial foraging based edge detection for cell image segmentation," *2015 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 3873–3876, 2015.
3. E. Tuba, A. Alihodzic, and M. Tuba, "Multilevel image thresholding using elephant herding optimization algorithm," *2017 14th Int. Conf. Eng. Mod. Electr. Syst. EMES 2017*, pp. 240–243, 2017.
4. U. Pal and B. B. Chaudhuri, "Machine-printed and hand-written text lines identi ® cation," *Pattern Recognit. Lett.*, vol. 22, pp. 431–441, 2001.
5. Deokate S., Uke N. (2018) Various Traditional and Nature Inspired Approaches Used in Image Preprocessing. In: Pawar P., Ronge B., Balasubramaniam R., Seshabhattar S. (eds) Techno-Societal 2016. ICATSA 2016. Springer, Cham.
6. R. Jayadevan, S. R. Kolhe, P. M. Patil, and U. Pal, "Database development and recognition of handwritten Devanagari legal amount words," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 304–308, 2011.
7. B. B. Chaudhuri and U. Pal, "A complete printed Bangla OCR system," *Pattern Recognit.*, vol. 31, no. 5, pp. 531–549, 1998.
8. K. C. Fan, Y. K. Wang, and T. R. Lay, "Marginal noise removal of document images," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 2001–Janua, pp. 317–321, 2001.
9. K. Ntirogiannis, B. Gatos, and I. Pratikakis, "Performance evaluation methodology for historical document image binarization," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 595–609, 2013.
10. G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. J. Perantonis, "A complete optical character recognition methodology for historical documents," *DAS 2008 - Proc. 8th IAPR Int. Work. Doc. Anal. Syst.*, pp. 525–532, 2008.
11. V. J. Dongre and V. H. Mankar, "Segmentation of Printed Devnagari Documents," in *Advances in Computing and Information Technology. Communications in Computer and Information Science*, 2011, vol. 198, no. May, pp. 211–218.
12. B. B. Chaudhuri and U. Pal, "An OCR System to Read Two Indian Language Scripts: Bangla and Devnagari (Hindi) B.," *Comput. Vision, Pattern Recognit. Unit*, pp. 1011–1015, 1997.
13. V. Bansal and S. R. M. K., "Segmentation of touching characters in Devanagari," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1655, pp. 151–156, 1999.
14. V. J. Dongre and V. H. Mankar, "Devnagari Document Segmentation Using Histogram Approach," *Int. J. Comput. Sci. Eng. Inf. Technoogy*, vol. 1, no. 3, pp. 46–53, 2011.
15. S. S. Shelke and S. S. Apte, "A Multistage Handwritten Marathi Compound Character Recognition Scheme using Neural Networks and Wavelet Features," *... Pattern Recognit.*, vol. 4, no. 1, pp. 81–94, 2011.
16. A. N. Holambe, S. N. Holambe, and R. C. Thool, "Comparative study of devanagari handwritten and printed character & numerals recognition using Nearest-Neighbor classifiers," *Proc. - 2010 3rd IEEE Int. Conf. Comput. Sci. Inf. Technol. ICCSIT 2010*, vol. 1, pp. 426–430, 2010.
17. V. J. Dongre and V. H. Mankar, "Devnagari Handwritten Numeral Recognition using Geometric Features and Statistical Combination Classifier," *Int. J. Comput. Sci. Eng.*, vol. 5, no. 10, pp. 856–863, 2013.
18. U. Pal, T. Wakabayashi, N. Sharma, and F. Kimura, "Handwritten numeral recognition of six popular Indian scripts," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 2, pp. 749–753, 2007.
19. S. Arora, D. Bhattacharjee, M. Nasipuri, D. K. Basu, and M. Kundu, "Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition," *Ind. Inf. Syst. 2008. ICIIS 2008. IEEE Reg. 10 Third Int. Conf.*, pp. 1–6, 2008.
20. S. Indolia, A. K. Goswami, S. P. Mishra, and P. Asopa, "Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach," *Procedia Comput. Sci.*, vol. 132, pp. 679–688, 2018.
21. F. Chollet and others, "Keras." 2015.
22. Martín Abadi et al, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." .
23. B. Chakraborty, B. Shaw, J. Aich, U. Bhattacharya and S. K. Parui, "Does Deeper Network Lead to Better Accuracy: A Case Study on Handwritten Devanagari Characters," *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, Vienna, 2018, pp. 411-416.

AUTHORS PROFILE

Dr. Nilesh J. Uke, He received the B.E. degree in Computer Science and Engineering the Amaravati Univeristy, India, in 1995, and the M.E. from Bharathi Vidhyapeeth in 2005 and Ph.D. degrees in Computer Science, from STUM University India, in 2014. He is currently a Principal at Trinity Academy of Engineering, Maharashtra, Pune. His current research interest includes Visual Computing, Artificial Intelligence, Human Computer Interface and Multimedia. He is member of IEEE, ACM and Life Member of the Indian Society for Technical Education (ISTE) and Computer Society of India (CSI).

.

Sarika T. Deokate, she is a research scholar at JJTU University Rajasthan. She received the BE in 2005 and ME degree in Computer Engineering from Pune University. She is currently an Assistant Professor. Her current research interest includes Image Processing, Document Analysis and Recognition. She is Life Member of the Indian Society for Technical Education (ISTE).