International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Demystification of Bilingual Optical Character Recognition System for Devanagari and English Scripts

Rohit Verma, Jahid Ali

Abstract: ***Research is deliberately going on in the field of pattern recognition. New ideas are developed and implemented in this field throughout the globe. Optical Character Recognition (OCR) is one of the inseparable applications of Pattern Recognition. Though extensive research is already reported in this field, but multilingual Optical Character Recognition is the most challenging aspect which is still, the need of the hour. Myriads of researchers are digging the information to gather the best solutions for the recognition purpose. In this research paper, we are purposing the steps for the recognition of Devanagari and English scripts simultaneously occurring in the documents. A new approach of segmentation and splitting the characters of both the scripts is also introduced for the benefits of researchers. Most commonly in the documents containing English and Devanagari scripts, English characters are already separated, the challenge is to separate the Devanagari characters. Algorithm to implement the challenging aspect to segment the Devanagari and Roman scripts simultaneously is also implemented in the present paper.***

Keywords : Optical Character Recognition, Segmentation, Feature extraction, Dataset, pattern recognition, image processing.

I. INTRODUCTION

Optical Character Recognition narrates the complete process of recognition of characters from the printed or hand written documents and process them for editable forms so that purposeful objectives could be achieved. The earliest research in the domain of Character Identification was reported in 1870, where Carey[1] developed an image transmission system termed as retina scanner. In the early stage the scope and applications of Character Recognition system was limited to hearing aids, processing of documents only[2], but with the advent of technology the spectrum become vast to a great extent and now a days researchers are inventing myriads of applications where character recognition algorithms can depict its hold. Though extensive research is reported with more than 95% accuracy [3] in the domain of character recognition. But frequency of research related to scripts of western countries are very high. Not much work is reported in the recognition of Indian scripts associating with roman characters scripts. Character recognition of documents containing multilingual scripts opens up a vast filed of research and challenges.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Rohit Verma*, Research Scholar, IKG Punjab Technical University, Jalandhar, India, rohitverma11@gmail.com

Dr. Jahid Ali, Director, Shri Sai Iqbal College of Management & IT, Pathankot, India, zahidsabri@rediff.com

It is noticed that twelve distinguish scripts [4] are used to write Indian Official languages. So great opportunities for the researchers to research in the field of multilingual character recognition system for the identification of Indian scripts associated with English scripts are available.

II. OCR PROCESS

The character recognition procedure is sliced into five steps: Digitization, Preprocessing, Segmentation, Features Extraction and Classification.

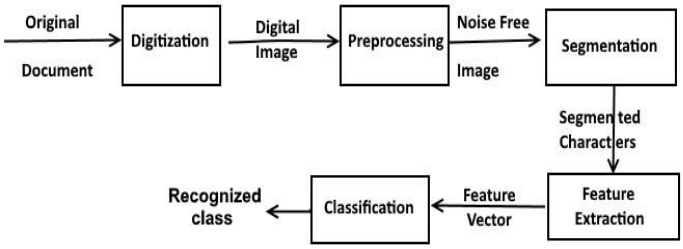


Fig 1: Block Diagram of OCR System.

A. Digitization

The output of the digitization process is the digital image of the document that can be employed for character recognition. A good quality scanner or a camera can be used under the presence of sufficient lightening conditions to generate the excellent results. But even after taking a lot of precautions, sometimes environmental conditions, device errors and other issues, the generated digital images are not fit for further analysis. Therefore, these images have to pass through the step of preprocessing.

B. Preprocessing

Preprocessing task is responsible for performing different roles. The digitized image received in the above step may suffer from several categories of noises. In the paper [4], R. Verma et al. describes different classes of noise that could suffer the performance of the OCR and several noise removal techniques that could be applied to achieve a noise free and clean document. Furthermore, the in the document the images are normalized. In our experiment it is assumed that the document is free from any kind of noise and comprises of printed Devanagari and English words.

C. Segmentation

The segmentation step is very critical and needs extensive care to achieve high degree of accuracy. Specially, in case of multilingual character recognition system more attention should be paid so that the segmented character could be properly separated for the next step of recognition. In the present work, firstly we have applied the process of skeletonization to achieve the given text in one pixel thin format. It is observed that region of almost every character of Devanagari script is overlapped with mattra (vowel form). So we have segmented the characters with mattra associated with that character.

Fig 2. Diagram to depict the process of segmentation

- *Proposed Algorithm for Segmentation*

1. Scan the printed text document and then binarize it.

2. Skeletonize the above scanned document to get one pixel format of each character and take negative of it.

3. Starting from the top horizontal row count the number of white pixels and find the row with maximum pixels.

4. Row with maximum white pixels is term as header line in Devanagari script.

5. Using vertical profile calculate pixel count N of every column of the image

If N ==1 then

set pixel value=0 in column location of row with maximum pixels.

End

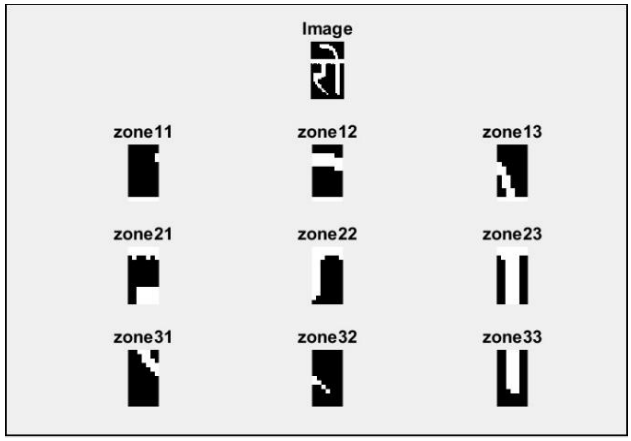
6. Take the negation of the split character image obtained.

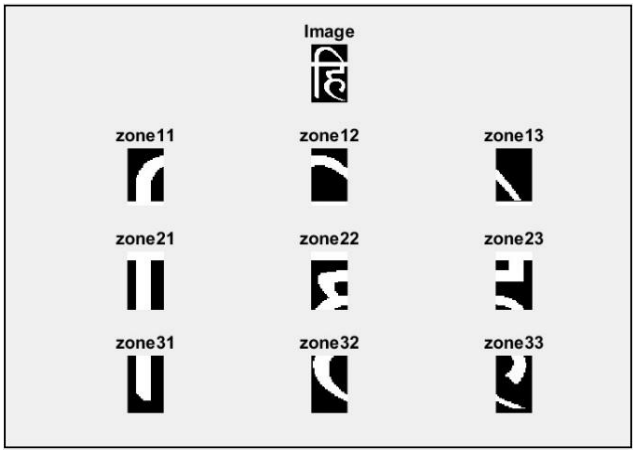
By the above proposed method of segmentation accuracy of more than 90% is achieved.

D. Feature Extraction

This is the extensively challenging part in the process of character recognition. This step contributes to prepare feature vector of significantly important features of the character to achieve high degree of accuracy. Generally those features are considered most vital which are translation, scale and orientation independent. These features can be shape oriented or non-shape based features. Shape oriented features include moment related features like Hu's Invariant moments, Native features include eccentricity, extent and different types of loops, Geometrical features include area calculations, geometrical distance calculations. In paper [5], Verma Rohit et al. describes most significant features employed for the identification of characters. The accuracy of the character recognition on the specified dataset was also described. In paper [6], Geetha Srikantan reported an accuracy of more than 99% in the recognition of numerals with gradient based contour encoding features. Negi et al. in the paper [7] reported an accuracy of 92% in the recognition of Telugu characters using template matching technique. Following are some of the features that could be considered for character recognition.

- Zone density features: Zone density features are shape based features. Here the segmented character is divided into nine zones of same size.
- Now the percentage zone density of pixels of each zone can be calculated by computing the total number of white pixels in each zone divide by the number of white pixels in the complete character.





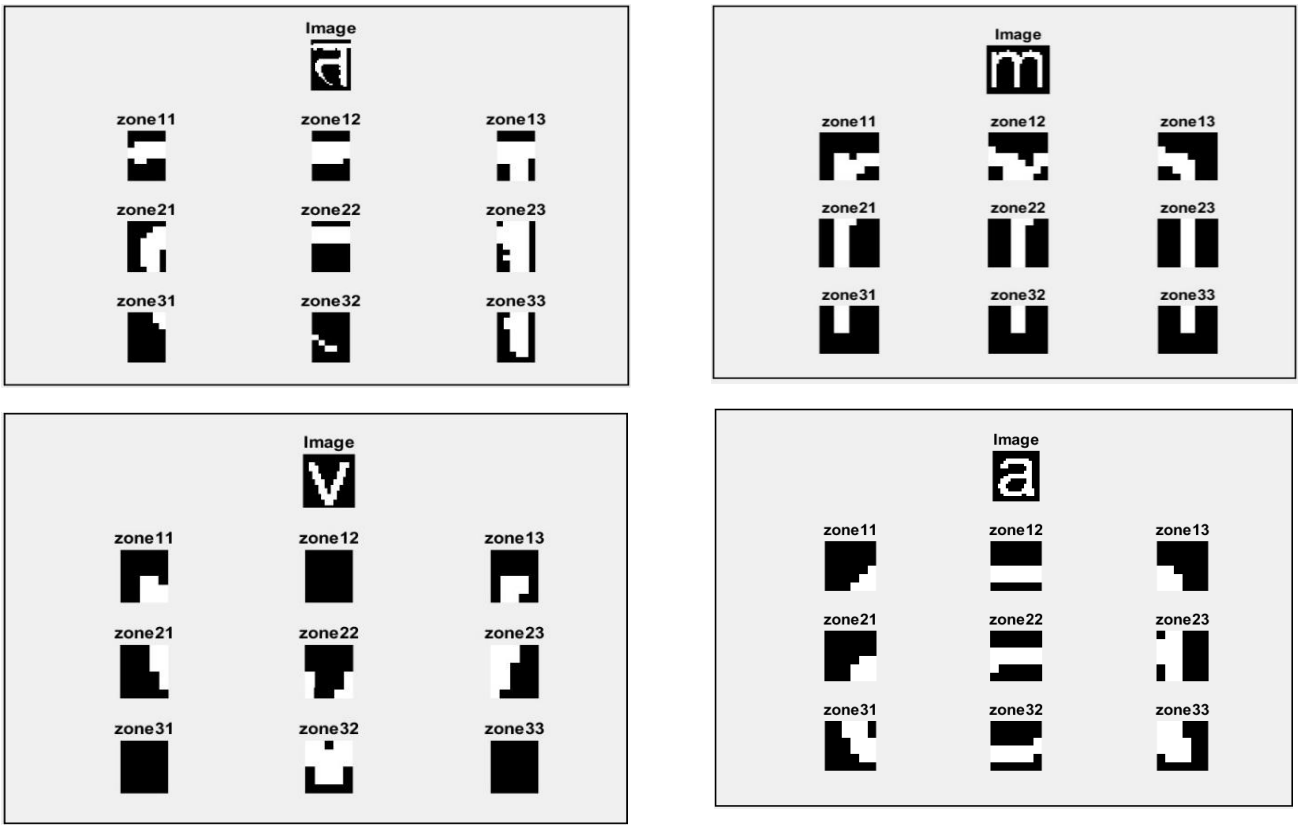


Fig 3. Zone wise segmentation of different segmented characters

Table 1. Percentage density of white pixels in each zone for different characters

	Z11	Z12	Z13	Z21	Z22	Z23	Z31	Z32	Z33
रो	0.0354	0.1102	0.0866	0.1457	0.1378	0.189	0.0551	0.0197	0.1142
हि	0.1271	0.0777	0.0553	0.1534	0.1602	0.1455	0.0956	0.0922	0.0709
त	0.0945	0.1144	0.1592	0.1194	0.0896	0.1791	0.0249	0.0199	0.1244
v	0.1053	0	0.1053	0.1053	0.0921	0.1711	0	0.2632	0
e	0.055	0.1193	0.0459	0.1835	0.1101	0.1101	0.1009	0.1284	0.1009
r	0.1538	0.2115	0.0769	0.2692	0.0577	0	0.1538	0.0769	0
m	0.1429	0.1825	0.119	0.119	0.119	0.1111	0.0635	0.0635	0.0635
a	0.0545	0.1091	0.0727	0.0727	0.1182	0.1364	0.1273	0.1091	0.1545

Results are generated with MATLAB 18.0

- Local Binary Pattern (LBP) features: Local Binary pattern is one of the robust method for recognition and classification of patterns. It is a powerful descriptor known for its computational simplicity and its tolerance against intensity values variations. This robust method labels the pixels of the image by thresholding the neighborhood of each pixel.

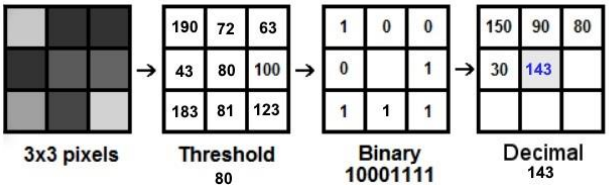
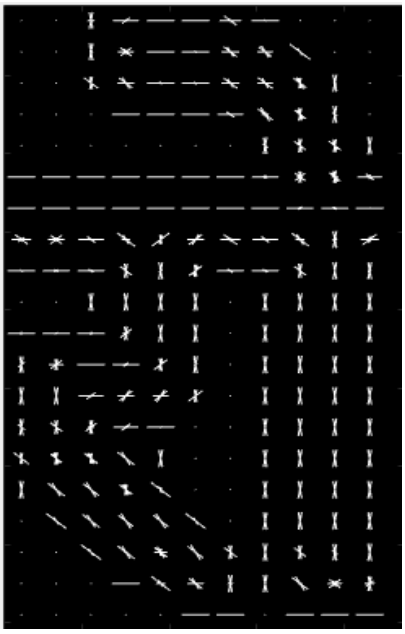


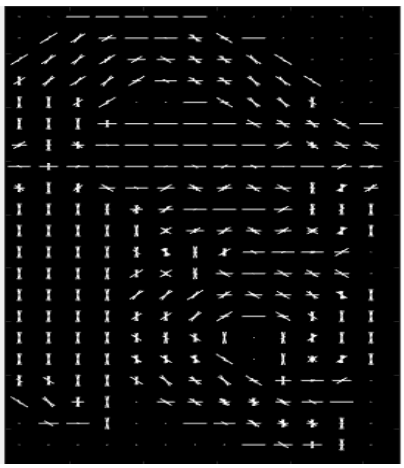
Fig 4. Block Diagram depicting the process of Local Binary Pattern

The neighboring pixels are tested against a threshold value. If their value is more than a certain pre-decided value, their value is turned to be "1" Otherwise zero value is assigned. In the paper [8], Hirwani et al. purposed a similar approach and the character to be recognized is splitted into 35 zones of equal size. The average of LBP features of each zone are then calculated and thus a total of 35 features were obtained. The similar approach was applied to a dataset of 284 characters and accuracy of 98.591 percent was reported.

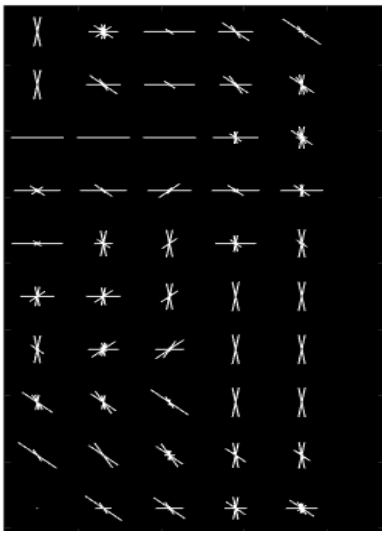
- Histogram of oriented gradient (HOG) Features: The HOG features can be used for object detection and recognition. The algorithm divides the image into tiny squared boxes and afterwards histogram of oriented gradients is computed in each box. Afterwards the output obtained is normalized using block-wise pattern and in each box the descriptor is returned for analysis. The box size could be [2 X 2], [4 X 4], or [8 X 8] etc. With increase in the cell size, one can loose small scale details. So cell size should be selected with great care to avoid information lose.



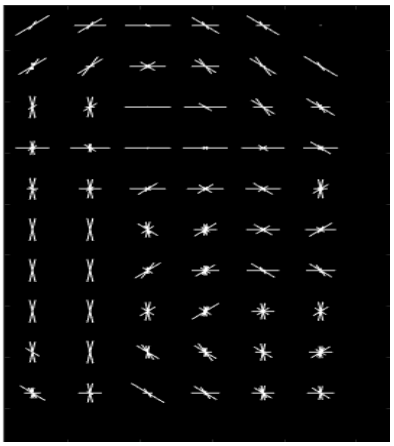
HOG of cell size 2X2 of image रो



HOG of cell size 2X2 of image हि



HOG of cell size 4X4 of image रो



HOG of cell size 4X4 of image हि

Fig 5. HOG representations of the tested character set

In the paper [9], Kamble et al. applied the same approach for the recognition of Marathi characters with FFANN and SVM classifiers and reported considerable accuracy for the recognition of character set under consideration.

- Hu's Invariant moments: The moment oriented features are pattern sensitive in the domain of recognition and classification. Here the distribution of various pixels relative to a fixed point is analyzed. As the distribution of massive pixels about a fixed point is invariant under translation, rotation and scaling. So desirable results of high degree of accuracy could be achieved by considering these features as a part of feature vector generated for the recognition of characters. Regular moment of order (p+q) can be represented by the moment function

$m_{pq} = \int \int x^p y^q f(x, y)\, dx\, dy$ where f(x,y) represents the image intensity function.

Hu in paper [10], introduced seven moments which are invariant under scaling, translation and rotation.

These moments are computed by using the following expressions.

$$I_1 = \eta_{20} + \eta_{02}$$

$$I_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$

$$I_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$

$$I_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$$

$$I_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$

$$I_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$

$$I_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2].$$

Here I_1 is similar to moment of inertia around the image's centroid where pixel intensities are similar to physical density. I_7 is skew invariant and has the capabilities of identifying images which are otherwise identical in shape.

- Other Features: In addition to above mentioned features myriads of potentially significant features are available that can be considered for the character recognition. With the complexity of calculations of these features no doubt the time of recognition will increase while training, but at the same time the performance of the OCR will dramatically increase manifold. Some of these features are:
- Extent: Extent is defined as the percentage area of the bounding box occupied by the character. It is computed by the number of white pixels occupied by the character divided by the total number of pixels present in the bounding box containing the character.
- Eccentricity: Eccentricity is defined as the deviation of the curve from the circulatory. In case of image it is the eccentricity of the ellipse having same second-moments as that of the region under consideration. It is measured as the ratio of distances between the foci to the length of major axis of the ellipse. Its value lies between zero and one. Ellipse having eccentricity one is a circle.
- Orientation**:** Orientation is another important feature. It denotes the angle between the major axis of the ellipse that has the same second-moments as that of region of interest, and the x-axis.

III. DATASET PREPARATION

This step is very time consuming and tedious. Feature extraction step is completely dependent on the process of dataset preparation. The degree of recognition of a character completely dependents on the size of the dataset. For the current research work font style of different Devanagari and English characters are observed from different sources and it is noticed that in most of the cases of Devanagari script most of the characters are overlapped vertically when they are associated with some mattra. Nearly 40 different font style characters are collected for the preparation of multilingual Devanagari and English dataset. Thus the dataset comprises of 14520 Devanagari and Roman characters. This dataset does not include "ardh akshra" and "Chandra bindu" of Devanagari Script. To prepare the dataset of these printed characters font size of 14 is considered. The results of any Optical Character recognition system can be expected from the size of the dataset used for training and testing phase. A dataset of considerably small size could not generate desirable results. Furthermore, dataset should be equipped with almost all possibilities of the style and font of each and every character of the domain under consideration. Since our area of interest is recognition of Devanagari and English script so we have considered only these two scripts.

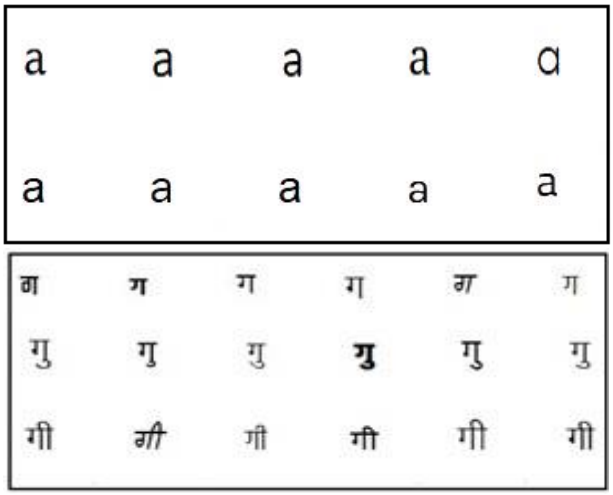


Fig 4. Part of the images from the dataset

IV. CLASSIFICATION

Classification is the machine learning engineering for assigning the labels to the unknown patterns on the basis of feature vector generated in the final and most critical step. Significant features converge towards the high degree of recognition accuracy. The output generated by the classifier is termed as prediction output. Countless classifiers are available that can be employed for the recognition purpose in numerous applications. But the challenge is to select the efficient classifier for a specific application. Some of the most intensively used classifiers are: K-nearest neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Convolutional Neural Network (CNN). The feature vector generated in the features extraction step can be applied as input to any one of these classifiers to achieve the desirable results. In the paper [11], S. Arora et al. applied SVM and ANN approaches to Devanagari character set comprises of 7154 characters and reported to receive accuracy to desirable level. J. Bakas in [12], investigates the recognition abilities of various classifiers. It is reported that MLP performed best with Oriya character set (95.20%) as well as Bengali character set (95.10%) , and SVM with radial basis function (RBF) kernel performs the best for Arabic character set (96.70%). After comparison it was reported that Knn performed better in comparison to other classifiers. Manish Kumar et al. in paper [13], reported an accuracy of 83% by employing FFANN on a dataset of 297 handwritten characters.

Reetika et al.[14], reported an accuracy of 98.77% in identifying English characters by experimenting SURF features with NN. Accuracy of 97.67% was reported by S.Roy [15] et al. by testing KNN classifier on MNIST numeral digits. In the paper [16], K. Gauri et al. achieved an accuracy of 97.16% for digits, 95.74% accuracy for capital letters and 92.19% for small letters by employing SVM classifier. A. El-Sawy [17], reported an accuracy of 94.9% by experimenting CNN on Arabic Characters. Thus from the literature it is reported that with potentially strong features and significantly important classification technique the degree of recognition could be taken to the higher levels.

V. RESULTS

The proposed segmentation technique employed for experiments with Matlab provide results more than 90% accurate for Roman and Devanagari characters simultaneously. Accuracy of more than 95% was reported with Devnagari script only.

VI. CONCLUSION

The main objective of this paper is to provide a straight-forward and detailed introduction for the new researchers to the functionality of OCR. A new approach of segmentation was also elaborated so that instead of segmenting the individual characters, they are segmented in association with the vowels or mattras associated with them. By using this approach the task of segmentation will no longer sound to be tedious and time consuming. Algorithm to segment the multi-script document would prove to be helpful to the new researchers working in the domain of pattern recognition.

In addition to this, concept of vital features like LBP, Zone density and HOG features are also elaborated for better understanding. The accuracy of recognition from the literature using these features enhances the importance of these features manifold. Finally, this paper describes various classification models that can be considered at the final stage for the classification of characters to attain high degree of accuracy. The accuracy attained by employing these classifiers by various researchers from the literature is also reported.

REFERENCES

1. J. Mantas, “An overview of character recognition methodologies,” *Pattern Recognit.*, vol. 19, no. 6, pp. 425–430, 1986.
2. A. Mir, A. Mir, S. A. Hannan, and Y. Perwej, “AN OVERVIEW AND APPLICATIONS OF OPTICAL CHARACTER RECOGNITION,” *Int. J. Adv. Res. Sci. Eng.*, vol. 8354, no. 3, pp. 261–274, 2014.
3. N. R. Soora and P. S. Deshpande, “Robust Feature Extraction Technique for License Plate Characters Recognition,” *IETE J. Res.*, vol. 61, no. 1, pp. 72–79, 2015.
4. R. Verma, “A Comparative Study of Various Types of Image Noise and Efficient Noise Removal Techniques,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 10, pp. 617–622, 2013.
5. R. Verma, “A Review of key features instrumental in recognizing the Multi-lingual Optical Characters,” *Int. J. Res. Eng. Appl. Manag.*, vol. Vol-05, no. Issue-02.
6. G. Srikantan, S. W. Lam, and S. N. Srihari, “GRADIENT BASED CONTOUR ECODING FOR CHARACTER RECOGNITION,” vol. 29, no. 7, 1996.
7. A. Negi, C. Bhagvati, and B. Krishna, “An OCR system for Telugu,” *IEEE*, pp. 1110–1114, 2001.
8. A. Hirwani, N. Verma, and S. Gonnade, “International Journal of Advanced Research in Computer Science and Software Engineering Efficient Handwritten Alphabet Recognition Using LBP based Feature Extraction and Nearest Neighbor Classifier,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 4, no. 11, pp. 549–553, 2014.
9. P. M. Kamble and R. S. Hegadi, “Handwritten Marathi character recognition using R-HOG Feature,” *Procedia - Procedia Comput. Sci.*, vol. 45, pp. 266–274, 2015.
10. M. K. Hu, “Visual Pattern Recognition by Moment Invariants,” pp. 66–70, 1962.
11. S. Arora, D. Bhattacharjee, M. Nasipuri, D. K. Basu, and M. Kundu, “Recognition of Non-Compound Handwritten Devnagari Characters using a Combination of MLP and Minimum Edit Distance,” *Int J. Comput. Sci.*, no. 4, pp. 107–120, 2010.
12. J. Bakas, “A Comparative Study of Various Classifiers for Character Recognition on Multi-script Databases,” *Int. J. Comput. Appl.*, vol.
13. M. K. Sahu, N. K. Dewangan, and M. T. Scholar, “Handwritten Character Recognition using Neural Network,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 6, no. 6, pp. 11–14, 2017.
14. R. Verma and R. Kaur, “Enhanced Character Recognition Using Surf Feature and Neural Network Technique,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 5565–5570, 2014.
15. S. Roy and M. Saravanan, “Handwritten Character Recognition using K-NN Classification Algorithm,” *IJARIIE*, no. 5, pp. 1245–1250, 2017.
16. G. Katiyar, A. Katiyar, and S. Mehfuz, “Off-Line Handwritten Character Recognition System Using Support Vector Machine,” *Am. J. Neural Networks Appl.*, vol. 3, no. 2, pp. 22–28, 2017.
17. A. El-Sawy, M. Loey, and H. El-bakry, “Arabic Handwritten Characters Recognition using Convolutional Neural Network,” *WSEAS Trans. Comput. Res.*, vol. 5, pp. 11–19, 2017.

AUTHOR PROFILE

Mr. Rohit Verma completed his Masters in Computer Applications in 2003. Presently he is working as Assistant Professor in School of IT at AIMETC, Jalandhar, Punjab, India. He has vast experience of teaching and research spread over reputed Institutes. He is specialized in Computer Graphics, Animation and software development. His domain of research is Pattern Recognition. He contributed in the field of character recognition by publishing research work towards character recognition.

Dr. Jahid Ali has vast research, teaching and administrative experience in SSGI Badhani, since 2002. He has specialized in Speech Recognition Technology, Artificial Intelligence, Advanced Data Structures, Applied Mathematical and Programming languages. Apart from many co-curricular and extra circular activities, he has a membership of NSS. He has published about 15 papers in National Journals of repute and guiding 4 Ph.D students from Punjab Technical University.