Offline Character Recognition Based on Structural Primitives

D.Sarvani, S.A.Bhavani, Tusar Kanti Mishra
Department of Computer Science and Engineering, ANITS
Visakhapatnam, India.

Abstract:- In latter years, character recognition is the most important topic in field of image processing. Character recognition still faces the problem of character matching. In this paper, we have conferred a complete Offline English character recognition system by using a suitable feature extraction algorithm that uses the structural feature primitives of the character image. To overcome this problem we have come up with a suitable feature extraction algorithm. The approach has the states like preprocessing, segmentation, feature extraction and classification and recognition. The offline character recognition system is tested on 26 each of the lowercase and uppercase English characters and 0-9 digits. We performed the experiments by taking different types of benchmark dataset images. The results so obtained are satisfactory. An overall rate of accuracy of 92.5% has been obtained. Performance comparison with other state-of-the-art methods is also carried out whereby the proposed method outperformed the others in terms of rate of accuracy.

***Keywords**:- Preprocessing; Segmentation; Feature Extraction; Classification; OCR.*

I. INTRODUCTION

Now a days, optical character recognition has been the interesting topic in the field of research. Optical Character Recognition (OCR) is the conversion of large amount of documents or images, either printed or handwritten into machine encoded text without any changes.OCR is one of the fields in pattern recognition, which converts an image of typewritten or handwritten, printed text into an text the computer can easily understand. OCR is used in the field of research in signal processing, artificial intelligence, pattern recognition, and machine vision.

In this world, there are different types of languages. We are using the English language because most of the people uses this language and can be easily understand to everyone. OCR uses the 26 lowercase letters i.e., from a to z, 26 uppercase letters i.e., A to Z and 0 to 9 numbers. Character recognition is of two types: offline and online and is shown in Fig.1. Offline character recognition, is a static recognition where the system scans and recognizes static images of the characters whereas, Online character recognition starts while the user is writing and performed at the same time. But, in this paper, we are completely discussing about Offline optical character recognition system based on English characters and Numerals.

OCR consists of phases like image acquisition, preprocessing, segmentation, feature extraction, classification and recognition. The task of preprocessing consists of image conversion and noise removal. In this paper, the most important state is Feature Extraction where OCR recognizes character based on the features .After the extraction of the features it will check the similarity between the given input characters to the inbuilt characters. Next phase is classification and recognition. In this phase we are using a template matching based on features method which gives the target output. The applications of OCR are mail sorting, invoice imaging, banking, health sector, digital libraries, automatic number plate recognition, handwritten recognition.

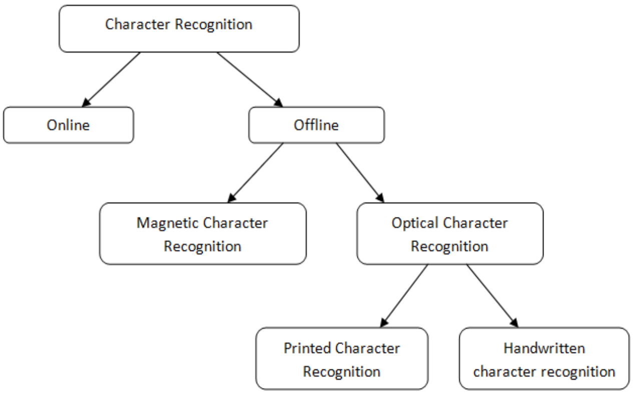

Character Recognition
Online
Offline
Magnetic Character Recognition
Optical Character Recognition
Printed Character Recognition
Handwritten character recognition

Fig 1:- Typical hierarchy of OCR system.

II. LITERATURE SURVEY

In [1], an experiment have been conducted on both handwritten and printed text with several feature extraction techniques like Gray Level Co-occurrence Matrix, Gabor filters, Distance profile, Zoning, Zernike moments, Projection histogram, Histograms of oriented Gradients and several classification techniques like Support vector machines, Naïve Bayes classifier, Artificial neural networks. Out of this , Multilayer Perceptron (MLP) in Artificial neural networks Classification with Histograms of oriented Gradients (HOG) feature gives better results on all of them with a good recognition rate. In [2], a review have been conducted on both online and offline character recognition for classification using a neural network algorithm and the results shown that the recognition rate and accuracy are higher for online character recognition than offline character recognition. In [3], they revised all the papers of handwritten English character recognition system and takes the different feature extraction and classification approaches and compared. Out of, all of this Fourier descriptor with phase and SVM classifier gives a better result than others. In [4], an experiment have been conducted on three template matching methods namely correlation methods, normalized cross correlation, and performance index method on character images of single or multiple words comparison. Out of all the three performance index gives best result. In [5], they conducted an experiment on English characters of hand-printed and they introduced a new feature extraction method that contains a segments by those segments they will identify the characters by using Fuzzy theory. In [6], an experiment have been conducted on offline English handwritten character recognition based on Convolutional Neural Networks with a modified LeNet-5 CNN model and improve the number of neurons in each layer and adding some error-correcting codes to the output. CNN is trained by using a error-samples-based reinforcement learning strategy. In [7], an experiment have been conducted in terms of Segmentation of cursive handwriting with segmentation processes like Word segmentation, Line segmentation, and Character segmentation. Out of the three processes they present a detail explanation on character segmentation. In Character Segmentation there are three types i.e., Implicit Segmentation, Explicit Segmentation and Holistic approaches. Out of these three Holistic approach works much better than the both. In [8], a Survey have been conducted on Printed English Character Recognition by using different feature extraction methods and different classification methods. But, they didn't explain which feature extraction and classification method works good. In [9], an experiment have been conducted on printed character recognition using different types of fonts like Arial, Times new Roman, Bookman old style, Tahoma by using Back propagation Neural Network with a feature extraction model. Out of all the fonts Times new Roman gives better results than other three. In [10], an experiment have been conducted on recognition of handwritten numerals of Odia language using hidden Markov model and recognition is performed by its class label.

In many a works, less importance has been given to the different versions and types of the fonts (Regular, italics, bold, bold-italics). However, much attention have been given to recognize the regular font types only. The time complexity of many a works is also at higher side. To overcome these challenges, an attempt has been made in this work which can address these issues properly.

The rest of this paper is organized as follows. The proposed work has been presented in section 3. This is followed by the experimental evaluation and result analysis and comparisons have been presented in section 4 which is followed by the concluding remarks in section 5.

III. PROPOSED WORK

The proposed system consists of 4 states: preprocessing, segmentation, feature extraction and classification and recognition shown in Fig.3. After acquiring an image the first state is preprocessing where the image gets converted into binary through binarization and removes the noise through morphological operations. The second state is segmentation where the text image gets segmented into individual characters. The third state is the crucial phase i.e., feature extraction. After the segmentation, through feature extraction it will extract the features. The final state is the classification and recognition where the template matching process will be done.

A. Image Acquisition

In this state, the system requires an input image i.e., a scanned image. The formats of an image is JPEG, PNG, BMP etc. This image is acquired through a digital camera, scanner or any other suitable digital input device.

B. Preprocessing

After the image acquisition, preprocessing is the first step in image processing and pattern recognition systems. Preprocessing is not a single step. It contains sequence of steps. In preprocessing state, the basic operation performed is Binarization. The input image is converted to gray scale image and then converted to binary image i.e., black and white which is called binarization shown in Fig.2. Most of the OCR works on Binarized images. In binarization there are different algorithms but we exploit sauvola technique.

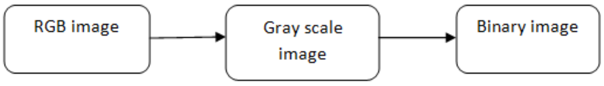


Fig 2:- Image conversion from RGB to Binary

- *Sauvola method*

It is a local thresholding technique useful for images where the background is not uniform, especially for text recognition. This technique provides a good results than other techniques and also removes the noise in background regions (non-text) and gives a better textual information. The local thresholding value is computed by the formula:

$$T(x,y) = m(x,y)*(1-k(1-s(x,y)/R))$$

Where T is threshold value for each pixel,
m is mean,
k is constant equal to 0.5,
s is standard deviation,
R is dynamic range of standard deviation s
(for grayscale images R=128).

- *Noise removal*

It means Removal of noise from an image. It is one of the biggest problem in optical character recognition process. In order to eliminate the noise we use morphology operations to detect and remove the noise present in the image.

Morphology operations are erosion, dilation, open and close.

Erosion: Erosion is the process of shrinking an image by deleting a layer of pixels to outer and inner boundaries of regions.

Dilation: Dilation is the process of growing an image by adding a layer of pixels to both outer and inner boundaries of regions.

- *Open:* Open operation is the combination of erosion followed by dilation.

- *Close:* Close operation is the combination of dilation followed by erosion.

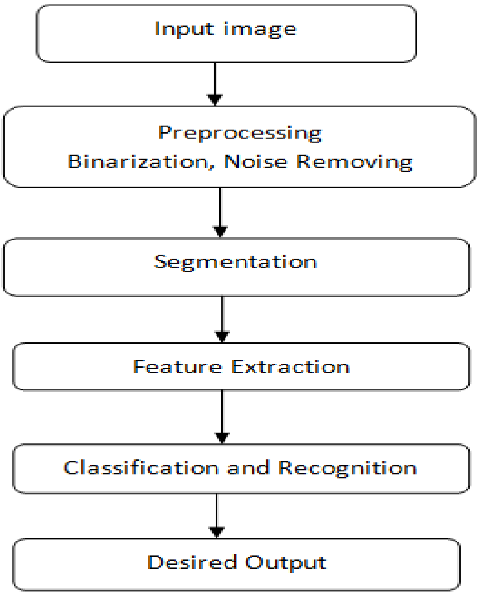


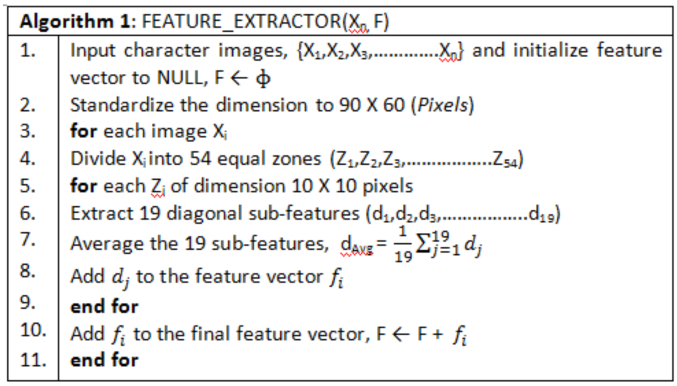
Fig 3:- Block diagram of proposed OCR system

C. Segmentation

Segmentation is one of the important stages of OCR process. In this, the input image gets divided into individual sequence of characters. The individual characters then goes into the feature extraction state and then it continues the process.

D. Feature Extraction

Feature extraction is the most important state for an OCR system. Features could be structural, topological and geometrical. Features are defined as significant definition of input character within less dimension space. The efficiency of features extracted from the character is defined as the recognition rate. The proposed algorithm has following steps:

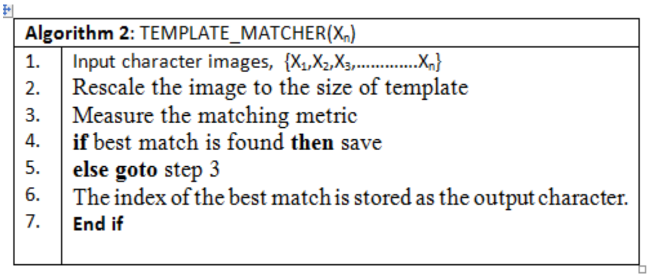


```
Algorithm 1: FEATURE_EXTRACTOR(Xn, F)
Input character images, {X1,X2,X3,............Xn} and initialize feature vector to NULL, F ← ϕ
Standardize the dimension to 90 X 60 (Pixels)
for each image Xi
    Divide Xi into 54 equal zones (Z1,Z2,Z3,................Z54)
    for each Zj of dimension 10 X 10 pixels
        Extract 19 diagonal sub-features (d1,d2,d3,................d19)
        Average the 19 sub-features, $d_{Avg} = \frac{1}{19}\sum_{j=1}^{19} d_j$
        Add $d_j$ to the feature vector $f_i$
    end for
    Add $f_i$ to the final feature vector, F ← F + $f_i$
end for
```

E. Classification and Recognition

The process of analyzing each character and designating it to the correct character class is called Classification. Classification uses the features extracted in the feature extraction state to identify the text in an image. In the classification, we are using a template matching technique based on features.

- *Template matching using features*



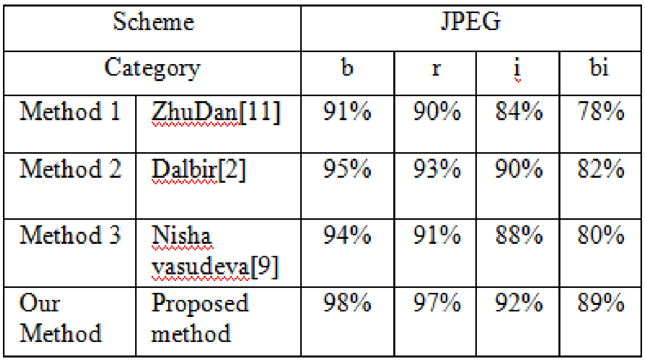
```
Algorithm 2: TEMPLATE_MATCHER(Xn)
1. Input character images, {X1,X2,X3,............Xn}
2. Rescale the image to the size of template
3. Measure the matching metric
4. if best match is found then save
5. else goto step 3
6.     The index of the best match is stored as the output character.
7. End if
```

IV. EXPERIMENTS AND RESULTS

The printed database contains 64 characters classes, collected from The Chars74K dataset. The database consists of 74K samples. The character samples are obtained from natural images is 7705, hand drawn characters samples is 3410 using a tablet PC, synthesised characters is 62992 from computer having different font styles like Bold(b) ,Regular(r), Italic(i) and Bold-Italic(bi) and different sizes. Therefore, this database consists of 64 characters classes with a total of 74K samples.

The experiments are conducted in MATLAB R2018 environment by using a Laptop with an operating system of windows 7 equipped with Intel(R) Core(TM) i3-4005U CPU 1.70GHz and 4GB RAM.

In classification state we have used a Calibri font family with different image formats like JPEG, PNG and BMP and different font styles for each format and style we have tested with a number of images and the results are represented in the Table.1,2 and 3 and the final comparison table with all the methods is represented in Table.4. The experiments have been conducted on database of printed characters. We found that the template matching algorithm provides a greater recognition rate and accuracy with an achievement of 92.5%.



Scheme		JPEG			
Category		b	r	i	bi
Method 1	ZhuDan[11]	91%	90%	84%	78%
Method 2	Dalbir[2]	95%	93%	90%	82%
Method 3	Nisha vasudeva[9]	94%	91%	88%	80%
Our Method	Proposed method	98%	97%	92%	89%

Table 1: Results of all the methods with jpeg image format

Scheme		JPEG			
Category		b	r	i	bi
Method 1	ZhuDan[11]	91%	90%	84%	78%
Method 2	Dalbir[2]	95%	93%	90%	82%
Method 3	Nisha vasudeva[9]	94%	91%	88%	80%
Our Method	Proposed method	98%	97%	92%	89%

Table 2. Results of all the methods with png image format

Scheme		BMP			
Category		b	r	i	bi
Method 1	ZhuDan[11]	87%	86%	75%	69%
Method 2	Dalbir[2]	92%	90%	83%	79%
Method 3	Nisha vasudeva[9]	90%	87%	82%	70%
Our Method	Proposed method	96%	94%	89%	85%

Table 3. Results of all the methods with bmp image format

Scheme		PNG			
Category		b	r	i	bi
Method 1	ZhuDan[11]	89%	88%	79%	72%
Method 2	Dalbir[2]	93%	91%	85%	78%
Method 3	Nisha vasudeva[9]	92%	89%	85%	75%
Our Method	Proposed method	97%	96%	90%	87%

Table 4. Comparison of accuracy and error percentage of different methods

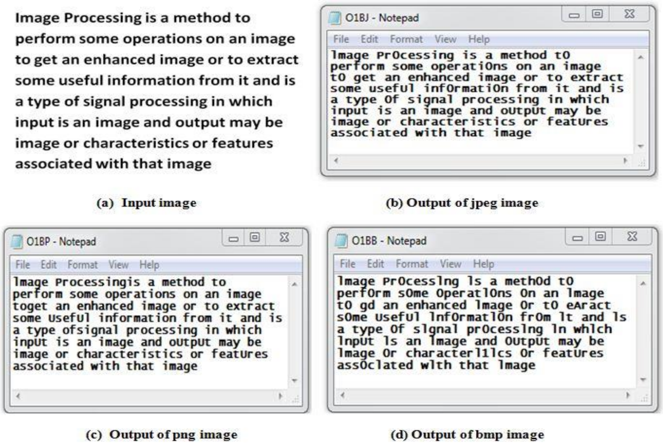


Image Processing is a method to
perform some operations on an image
to get an enhanced image or to extract
some useful information from it and is
a type of signal processing in which
input is an image and output may be
image or characteristics or features
associated with that image

(a) Input image

```
lmage PrOcessing is a method tO
perform some operatiOns on an image
tO get an enhanced image or to extract
some useful infOrmatiOn from it and is
a type Of signal processing in which
input is an image and oUtpUt may be
image or characteristics or featUres
associated with that image
```

(b) Output of jpeg image

```
lmage Processingis a method to
perform some operations on an image
toget an enhanced image or to extract
some UsefUl information from it and is
a type ofsignal processing in which
inpUt is an image and oUtpUt may be
image or characteristics or featUres
associated with that image
```

(c) Output of png image

```
lmage PrOcesslng ls a methOd tO
perfOrm sOme OperatlOns On an lmage
tO gd an enhanced lmage Or tO eAract
sOme UsefUl lnfOrmatlOn frOm lt and ls
a type Of slgnal prOcesslng ln whlch
lnpUt ls an lmage and OUtpUt may be
lmage Or characterl1lcs Or featUres
assOclated wlth that lmage
```

(d) Output of bmp image

Fig 4:- Input and output images of bold font with different image formats

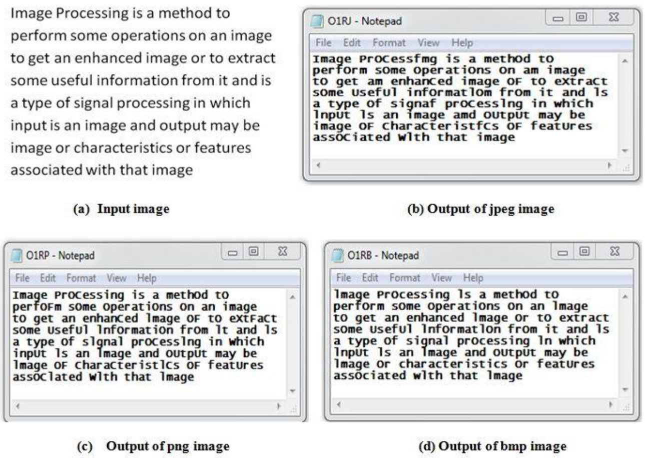


Image Processing is a method to
perform some operations on an image
to get an enhanced image or to extract
some useful information from it and is
a type of signal processing in which
input is an image and output may be
image or characteristics or features
associated with that image

(a) Input image

```
Image PrOcessfmg is a methOd to
perfOrm sOme OperatiOns On am image
to get am enhanCed image OF to eXtraCt
sOme UseFUl informatlOm frOm it and ls
a type Of signaf prOcesslng in which
lnpUt ls an image amd OUtpUt may be
image OF CharaCteristfCs OF featUres
assOCiated Wlth that image
```

(b) Output of jpeg image

```
Image PrOCessing is a methOd tO
perfOFm sOme OperatiOns On an image
tO get an enhanCed lmage OF to eXtFaCt
sOme UsefUl lnformatiOn frOm lt and ls
a type Of slgnal prOcesslng in which
inpUt ls an lmage and OUtpUt may be
lmage OF CharaCteristlCs OF featUres
assOClated Wlth that lmage
```

(c) Output of png image

```
lmage PrOcessing ls a methOd tO
perfOrm sOme OperatiOns On an lmage
tO get an enhanced lmage Or to extract
sOme UsefUl lnformatlOn frOm it and ls
a type Of signal prOcessing ln which
lnpUt ls an lmage and OUtpUt may be
lmage Or characteristics Or featUres
assOciated wlth that lmage
```

(d) Output of bmp image

Fig 5:- Input and output images of regular font with different image formats

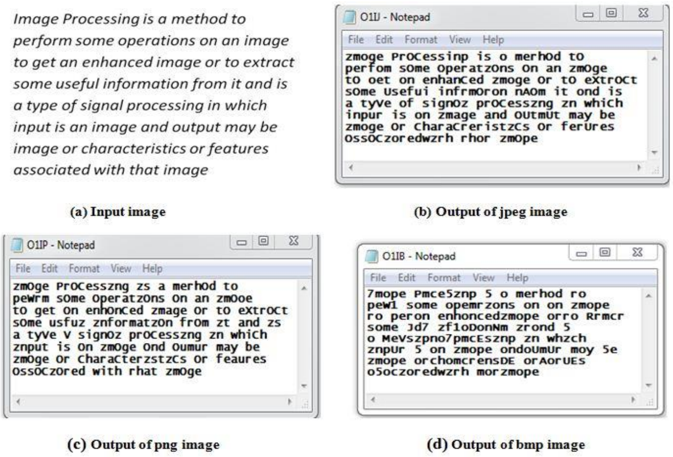


Image Processing is a method to
perform some operations on an image
to get an enhanced image or to extract
some useful information from it and is
a type of signal processing in which
input is an image and output may be
image or characteristics or features
associated with that image

(a) Input image

```
zmoge PrOCessinp is o merhOd to
perfom sOme OperatzOns On an zmOge
tO oet on enhanCed zmoge Or to eXtrOCt
sOme Usefui infrmOron nAOm it ond is
a tyVe of signOz prOCesszng zn whiCh
inpur is on zmage and OUtmUt may be
zmoge Or CharaCreristzCs Or ferUres
OssOCzoredwzrh rhor zmOpe
```

(b) Output of jpeg image

```
zmOge PrOCesszng zs a merhOd to
peWrm sOme OperatzOns On an zmOoe
tO get On enhOnCed zmage Or tO eXtrOCt
sOme usfuz znformatzOn frOm zt and zs
a tyVe V signOz prOCesszng zn whiCh
znput is On zmOge Ond Oumur may be
zmOge Or CharaCterzstzCs Or feaures
OssOCzOred with rhat zmOge
```

(c) Output of png image

```
7mope Pmce5znp 5 o merhod ro
pew1 some opemrzons on on zmope
ro peron enhoncedzmope orro Rrmcr
some Jd7 zf1oDonNm zrond 5
o MeVszpno7pmcEsznp zn whzch
znpUr 5 on zmope ondoUmUr moy 5e
zmope orchomcrensDE orAorUEs
o5oczoredwzrh morzmope
```

(d) Output of bmp image

Fig 6:- Input and output images of italic font with different image formats

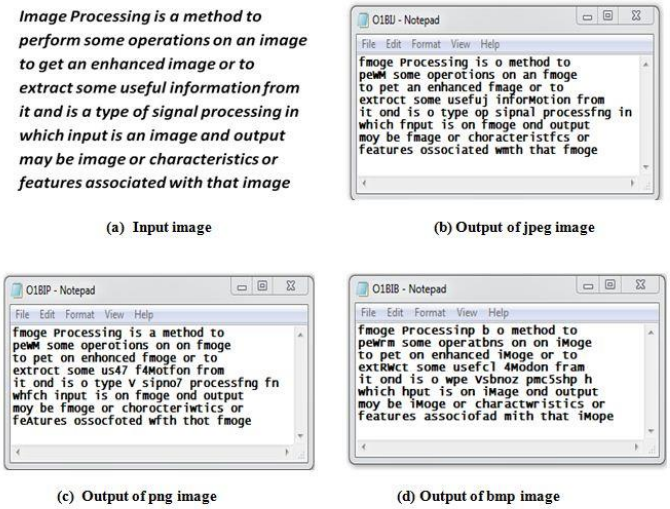


Image Processing is a method to
perform some operations on an image
to get an enhanced image or to
extract some useful information from
it and is a type of signal processing in
which input is an image and output
may be image or characteristics or
features associated with that image

(a) Input image

```
fmoge Processing is o method to
peWM some operotions on an fmoge
to pet an enhanced fmage or to
extroct some usefuj inforMotion from
it ond is o type op sipnal processfng in
which fnput is on fmoge ond output
moy be fmage or choracteristfcs or
features ossociated wmth that fmoge
```

(b) Output of jpeg image

```
fmoge Processing is a method to
peWM some operotions on on fmoge
to pet on enhonced fmoge or to
extroct some us47 f4Motfon from
it ond is o type V sipno7 processfng fn
whfch input is on fmoge ond output
moy be fmoge or chorocteriwtics or
feAtures ossocfoted wfth thot fmoge
```

(c) Output of png image

```
fmoge Processinp b o method to
peWrm some operatbns on on iMoge
to pet on enhanced iMoge or to
extRWct some usefcl 4Modon fram
it ond is o wpe Vsbnoz pmc5shp h
which hput is on iMage ond output
moy be iMoge or charactwristics or
features associofad mith that iMope
```

(d) Output of bmp image

Fig 7:- Input and output images of bold-italic font with different image formats

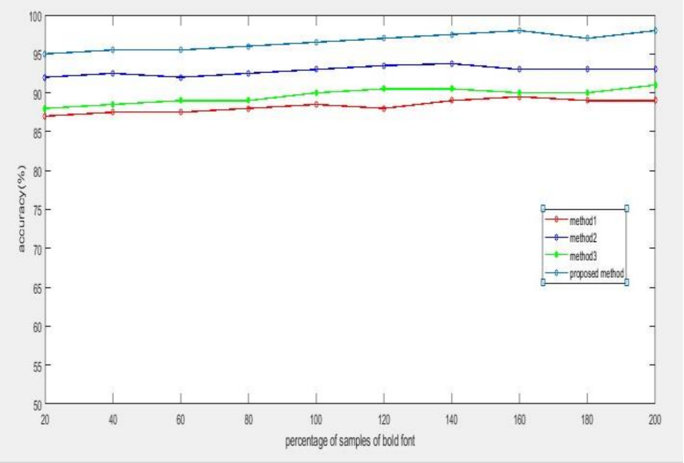


Fig 8:- Comparison of different methods with bold font of jpeg image format

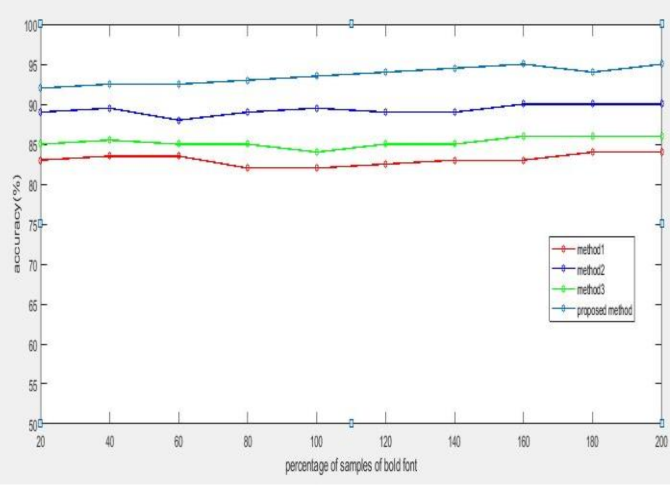


Fig 9:- Comparison of different methods with bold font of png image format

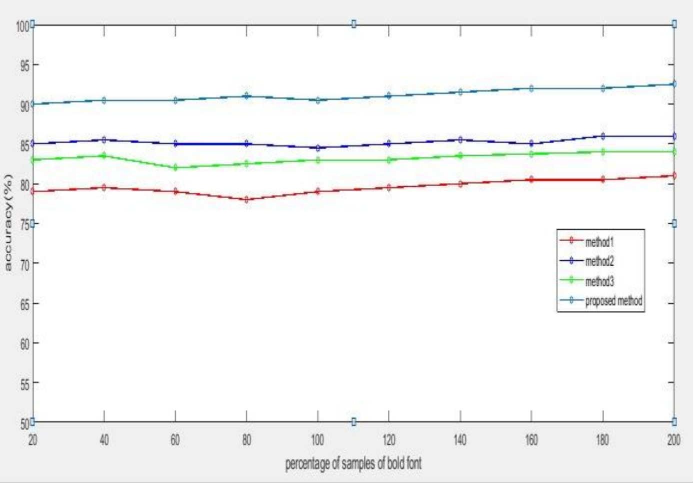


Fig 10:- Comparison of different methods with bold font of bmp image format

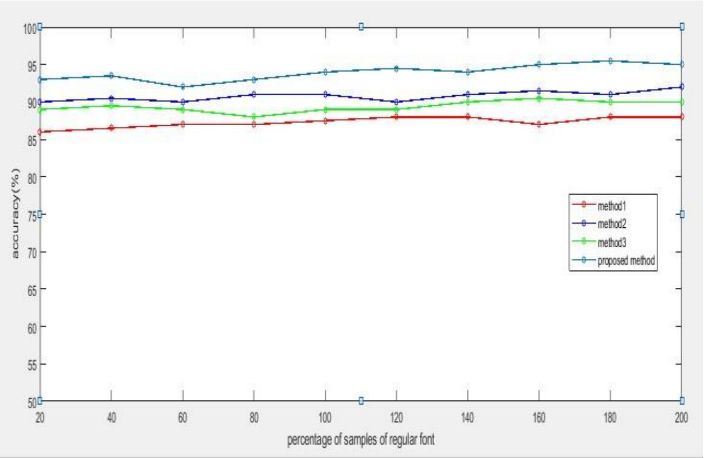


Fig 11:- Comparison of different methods with regular font of jpeg image format

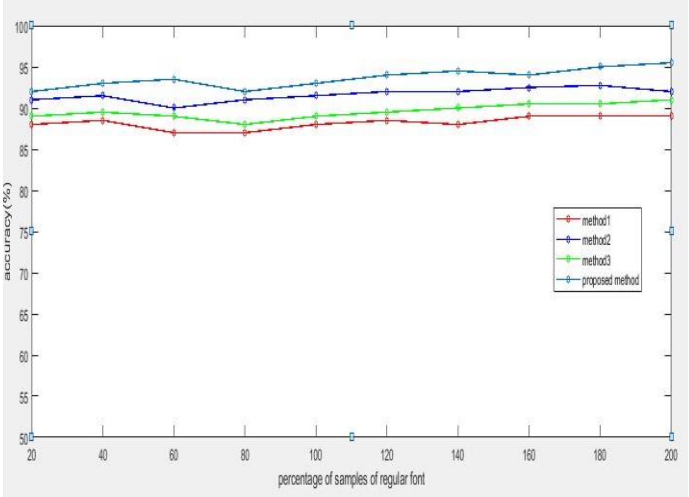


Fig 12:- Comparison of different methods with regular font of png image format

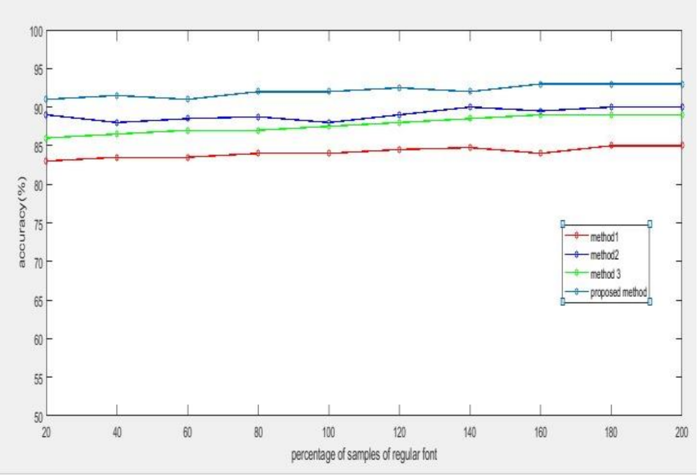


Fig 13:- Comparison of different methods with regular font of bmp image format

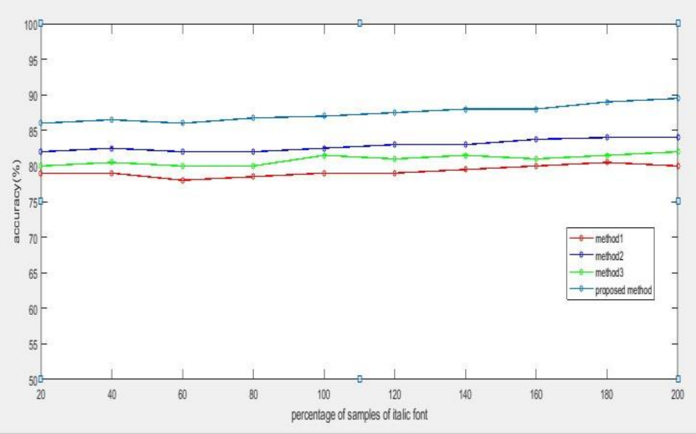


Fig 14:- Comparison of different methods with italic font of jpeg image format

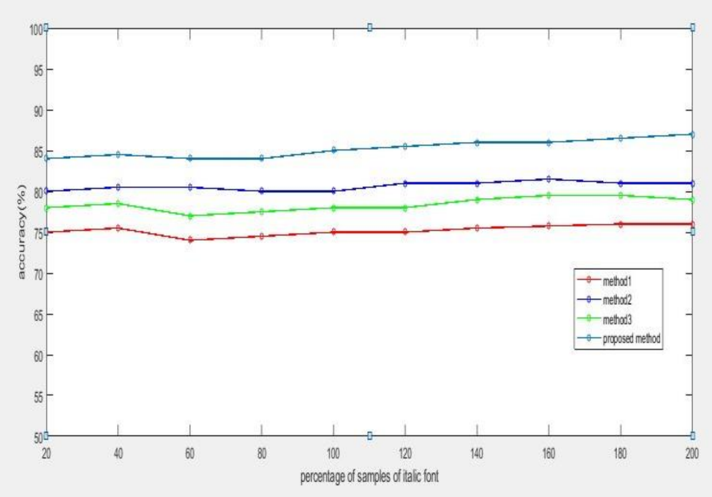


Fig 15:- Comparison of different methods with italic font of png image format

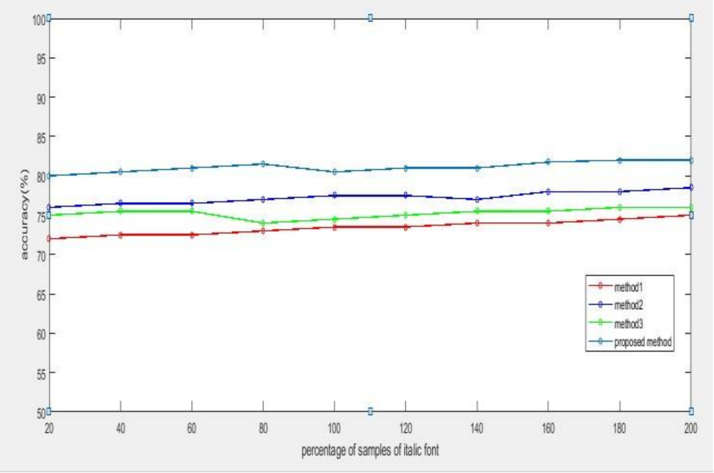


Fig 16:- Comparison of different methods with italic font of bmp image format

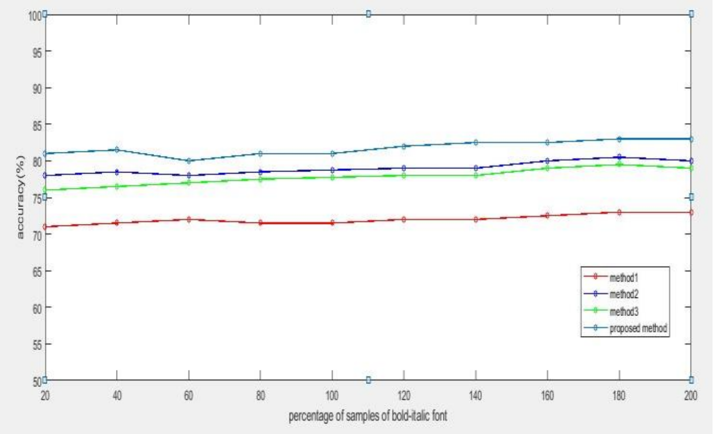


Fig 17:- Comparison of different methods with bold-italic font of jpeg image format

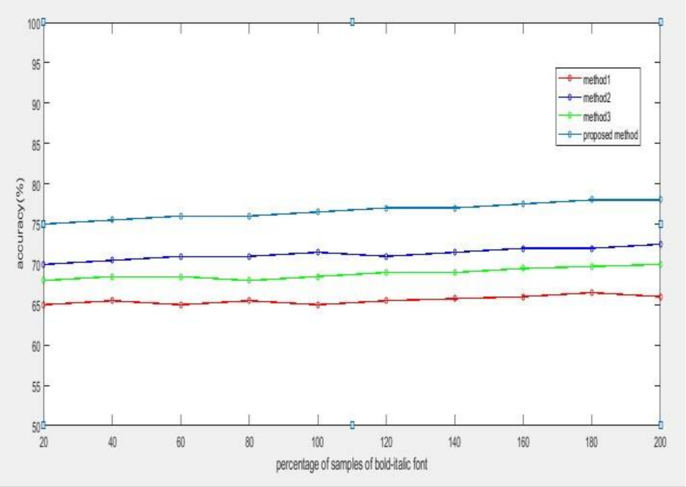


Fig 18:- Comparison of different methods with bold-italic font of png image format

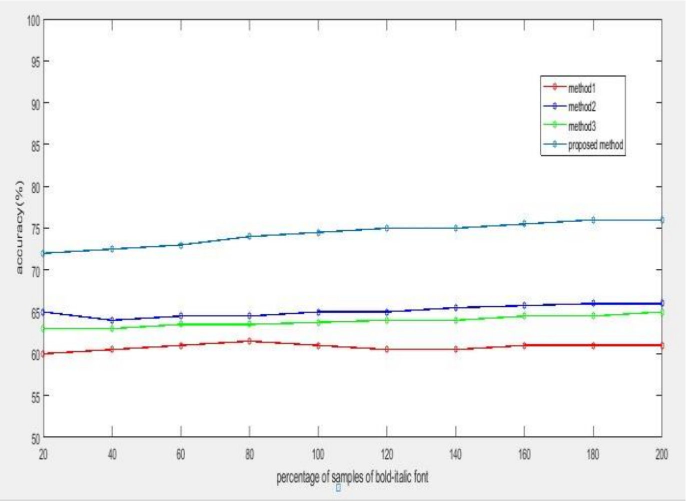


Fig 19:- Comparison of different methods with bold-italic font of bmp image format

V. CONCLUSION

In this paper, an Offline English character recognition system for different types of image formats and different fonts with a suitable feature extraction method. The experiments were conducted on 74K dataset using different feature extraction methods with a Calibri font family. From the tests the results obtained are shown in the Table.4 and is identified that our proposed method gives better results and highest accuracy than other methods.

REFERENCES

[1]. H. el bahi, Z. Mahani, A. Zatni and S. Saoud, "A robust system for printed and handwritten character recognition of images obtained by camera phone," wseas transactions on signal processing, volume 11, 2015.

[2]. Dalbir, Sanjiv kumar singh, "Review of online & offline character recognition," international journal of engineering and computer science issn:2319-7242,volume 4 issue 5 may 2015.

[3]. Nisha sharma, Tushar patnaik, Bhupendra kumar, "Recognition for handwritten english letters: a review ," international journal of engineering and innovative technology (ijeit) ,volume 2, issue 7, january 2013.

[4]. Dr.S.Vijayarani , Ms. A.Sakila, "Template matching technique for searching words in document images," international journal on cybernetics & informatics (ijci) vol. 4, no. 6, december 2015.

[5]. Puttipong Mahasukhon, Hossein Mousavinezhad and Jeong-young song, "Hand-printed english character recognition based on fuzzy theory," ieee,2012.

[6]. Aiquan yuan, Gang bai, Lijing jiao and Yajie liu, "Offline handwritten english character recognition based on convolutional neural network," IAPR international workshop on document analysis systems,2012.

[7]. Amandeep kaur, Seema baghla and Sunil kumar, "Study of various character segmentation techniques for handwritten off-line cursive words: a review," international journal of advances in science engineering and technology,issn:2321-9009, volume- 3, issue-3, july-2015.

[8]. Sagar s. Dutt and Prof. jay d. Amin , "Printed english character recognition using feature based matching and error correction: a survey," international journal of innovative research in technology, volume 2 issue 7,issn: 2349-6002, december 2015.

[9]. Nisha vasudeva, Hem jyotsana parashar and Singh vijendra, "Offline character recognition system using artificial neural network," international journal of machine learning and computing, vol. 2, no. 4, august 2012.

[10]. Tusar kanti mishra , Banshidhar majhi, Pankaj k sa and Sandeep panda, "Model based odia numeral recognition using fuzzy aggregated features," higher education press and springer-verlag berlin heidelberg ,2014.

[11]. Z. Dan and C. Xu, "The recognition of handwritten digits based on bp neural network and the implementation on android," third international conference on intelligent system design and engineering applications, pp. 1498–1501, 2013.

[12]. Ramesh Kumar Mohapatra, Tusar Kanti Mishra, Sandeep Panda and Banshidhar Majhi, "OHCS: A database for handwritten atomic Odia Character Recognition," IEEE,2015.

[13]. Tusar Kanti Mishra, Banshidhar Majhi and Sandeep Panda,"A comparative analysis of image transformations for handwritten Odia numeral recognition," IEEE,2013.

[14]. Z. Zhao, C.L. Liu and M. Zhao. "Handwriting representation and recognition through a sparse projection and low-rank recovery framework," international joint conference on neural networks (ijcnn), pp. 1-8, 2013.

[15]. G. Mayraz and G. E. Hinton, "Recognizing handwritten digits using hierarchical products of experts" eee transactions on pattern analysis and machine intelligence, vol 24, pp. 189–197,2002.