Received: December 5, 2017

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Identification of Broken Characters in Degraded Documents

Kaviya Ramalingam[1]* **Ramamurthy Bhojan[1]**

[1]*Department of Computer Science, Christ University, Hosur Road, Bengaluru, Karnataka, India*
* Corresponding author's Email: kaviya.r@cs.christuniversity.in

Abstract: Optical Character Recognition (OCR) deals with the recognition of characters in a text document. Steps like Preprocessing, Segmentation and Recognition are embedded in the OCR machine. When a document is scanned it will be taken into OCR and will recognize the characters. But noisy scanning of documents, low-quality printed documents and thresholding error leads to the generation of broken characters. When these documents are given as inputs into OCR, the recognition becomes a tedious process since the broken characters are misunderstood by the OCR machine. So the broken characters have to be identified and segmented separately. This work aims to enhance the degraded documents with broken characters using image processing techniques. For identifying or recognizing the broken character from the image various techniques like vertical projection profile, horizontal projection profile, chain code, mean based thresholding are used. The lines from the document are separated using line segmentation. Separate characters are extracted using Vertical Projection Profile and Horizontal Projection Profile. The character is identified using chain coding. The broken characters are found from them using Mean-based Thresholding and is merged using Heuristic information. The proposed method achieves an accuracy of 92.88% and also performs well for color image documents as well as black and white image documents also because of the effective preprocessing.

Keywords: OCR, Horizontal projection profile, Vertical projection profile, Chain code, Mean-based thresholding, Heuristic information.

1. Introduction

This Digital conversion of an Image to perform some operations on it to enhance or to extract some information from the image is known as Image Processing. Optical Character Recognition is a technique that is used to identify the characters in the images of text document and make them into an editable text in which searching and editing can be done easily[1, 2]. This has helped the humans in transferring the contents from a paper into digitized format without much human work. The result is much accurate if the document is clear. It is used in various fields such as banking to recognize the numbers in cheque, legal documents and health sectors to make it digitized so that there is less paperwork to be done. It is also used in the field of education, finance and government agencies [3, 4]. It makes the data collection and analysis work simple and easy. But in most of the old documents or in historical records the scripts gets worse as it ages. Also poor quality of scanner and photocopies makes worse documents with broken characters [5]. Damaged characters are the ones that is broken vertically or horizontally into two halves or even more. These damaged characters are a big challenge to OCR systems as they could not recognize them properly [6]. For overcoming this problem, a method is proposed that will segment the characters from the image and identify the broken characters in that and will rebuild it. This will finally give an enhanced document having a good quality so that error rate of OCR is reduced. For this purpose an approach called Mean based thresholding with chain coding is proposed which will solve the problem of wrong recognition rates in OCR machine that achieves considerable results. This paper is organized in a way that the works related to the proposed methods are described in section 2, proposed method is explained in detail in section 3,

Implementation methodology is given in section 4, Results and discussion are presented in section 5, and Conclusion is given in section 6.

2. Related work

There are many proposed methods for various regional languages to enhance the broken characters of the document. A literature survey of the existing models and methods is given in this section.

In this paper the Gujarati characters are segmented and recognized using techniques like Otsu's method, vertical projection profile and horizontal projection profile. In vertical projection profile the characters are identified column-wise and in horizontal projection profile the characters are identified row-wise. A MBT approach is used to identify the broken characters in which the characters with the width above the mean value are considered complete and those that are below the mean value are said to be broken. After the identification of broken characters they are merged by using space Information present in the database. An efficiency of 79.93% is achieved in this proposed method. But this method works only with the plain document images and not with the graphics image [7].

In this paper firstly the input image is saved as bitmap image and then its threshold value is set to 200. The bitmap image is segmented into individual characters, called segments by applying the dynamic profile projection technique. This paper proposed neighboring pixel and dynamic projection profile technique to identify the broken as well as multiple touching characters in Gurumukhi script. It has obtained an efficiency of 95.9%. Furthermore this can be extended to recognize the overlapping characters. However for skewed documents this doesn't fit well when it searches for neighbouring pixels. [8].

This paper proposed an end detection approach to recognize the damaged characters. The report that is degraded is processed to take out the noises in it. Then the photo is transformed to binary image. The binary image is then segmented by region based segmentation methods. The characters that are broken are isolated and thinned to get the border of it. This is then reconstructed with the aid of filling the gaps of the damaged character using end point algorithm proposed on this paper. However it could not attain an efficient result as the reconstructed image does not fill the gaps properly. [9].

This paper focus on the enhancement of broken and degraded kannada characters by using end detection algorithm. First the characters are identified using the region based segmentation that is Vertical Projection Profile and Horizontal Projection Profile. After this, the detected character is normalized and thinned to get the edge points. After detecting the edge points the line segment is drawn using connected components and is checked whether the character is broken or not. If it is broken then it is joined by using polynomial curve fitting. This proposed work gives 89% of accuracy. However the method is experimented on database of broken characters separately. It has not been implemented on degraded documents [10].

In this paper the author proposed a neighboring pixel and end point detection method to identify the damaged characters or the overlapping characters and segment it to rebuild it. They first segment the isolated characters broken characters and overlapped characters. This method works well for documents with broken or overlapped characters but is not suitable for documents containing skewed lines. But the inputs taken for this document is all handwritten images that are very clear to be processed. So this does not work well with degraded documents as the preprocessing techniques does not match the required criteria [11].

This paper proposed a neural network based approach for identifying the broken characters in damaged documents. The image is binarized by applying a Gaussian blur and grids are formed on the binarized image based on the area occupied by that character. Then broken and correct characters are identified by taking features from the character and giving it as an input into Multi Layer Feed Forward Neural Network Classifier (MLFNC). This method achieved an accuracy of 68.33%. This author has also tested the method with single characters and not as a document so when it comes to degraded documents or skewed documents this method does not work well [12].

This paper proposed a method to identify the broken or damaged characters because OCR machine interprets them as a wrong character. This paper proposed an approach to find the broken pieces from the isolated characters using an optimal set partition method. This approach is tested in various American and thai document and the results obtained were upto the expectation which made OCR to recognize the characters correctly. However the proposed method does not suit well for documents with high intensity graphics and skewed images [13].

In this paper, only vertically broken characters are considered. The horizontally broken characters problem was already solved in some other papers. In

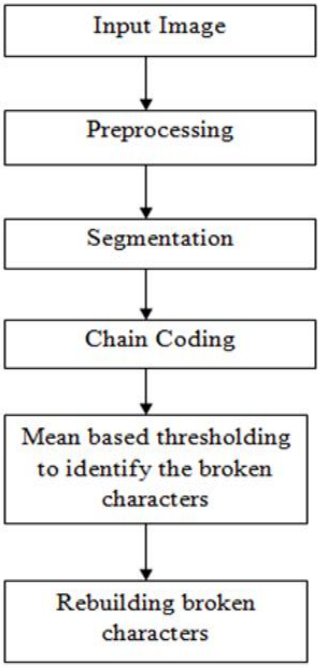

Figure.1 Flow chart for proposed method

this paper two techniques are used for identifying broken characters namely overlapping area and character code. Once the broken characters are identified then it is rebuilt using connection by considering both the overlapping areas and character codes. It was implemented in Visual basic studio and had good recognition result. Sometimes it makes erroneous recognition also, which was the drawback of this method [14].

3. Proposed method

In this proposed work the following Fig. 1 shows the flow chart of the steps that are to be followed to enhance the damaged documents with broken characters.

Input Image – The degraded document that is to be enhanced to rectify the broken characters is given as input by scanning it through a scanner. A sample image is given as input which is shown in Fig. 1.

Pre-processing – The image which is taken as input is a degraded document containing broken characters. This image will be having a noise that makes the OCR to erroneously recognize the characters [15]. This is rectified by applying some preprocessing techniques to smoothen or sharpen the image and to remove the noise in it [16]. At first the image that is taken is converted into black and white using preprocessing methods and then only segmentation is carried out [17].

Segmentation – In this process the image is segmented into separate grid or components based on the information present in the degraded document. To make this segmentation useful region based segmentation techniques like vertical and horizontal projection profile techniques are used.

Horizontal projection profile is used to segment the individual lines from a text document. A threshold value is set based on the intensity level which is marked from the histogram of the result of horizontal projection profile.

In vertical projection profile the individual characters are separated from the result of horizontal projection profile based on the threshold value taken from the histogram of the individual characters separated.

Chain Coding – Chain coding is a method that is used to identify the shape of the character which are segmented from horizontal and vertical projection profile. It will produce a several outputs out of which the top-quality and the closer one may be chosen [18].

Mean based thresholding – This method is used to identify the broken character and full character. If the character matches 95% and more of the mean value then it is classified as full character else it is considered as broken character.

Rebuilding – Once the broken characters are identified they are reconstructed using the heuristic information present in the database.

4. Implementation

Matlab is considered to be one of the good tools for image processing and so the implementation is done using Matlab.

Step 1 – Scanned text document is given as input into the Matlab software by using imread() function which is an inbuilt function to read the image in Matlab. The following Fig. 2 shows one sample input document image.

Step 2 – Preprocessing is a process in which the damaged document given as input is processed to take out the noise in that image [19]. Preprocessing method OTSU is used to make the image into binary image [20]. The resultant image will be a bianary image without noise that is shown in Fig. 3.

Step 3 – Characters are segmented from the document using region based segmentation techniques. The document is scanned both horizontally and vertically to isolate a separate character [21].

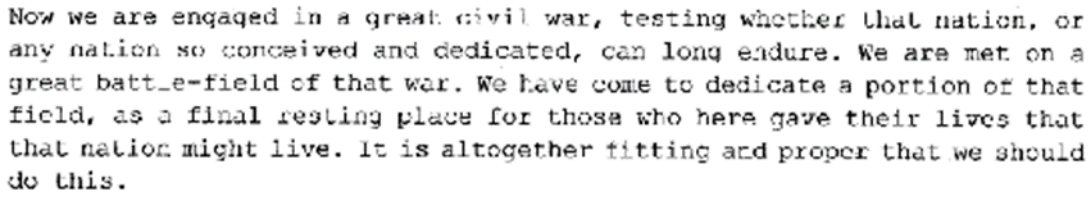

Figure.2 Input image

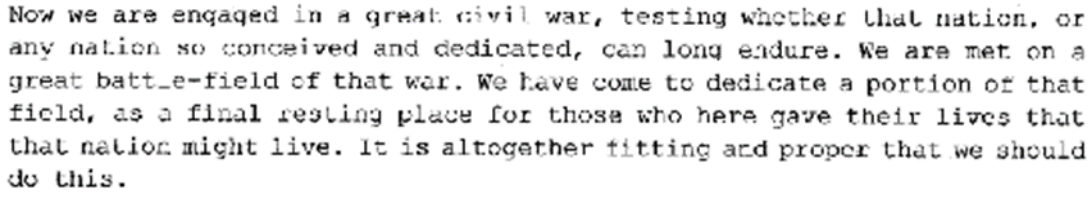

Figure. 3 Input image without noise

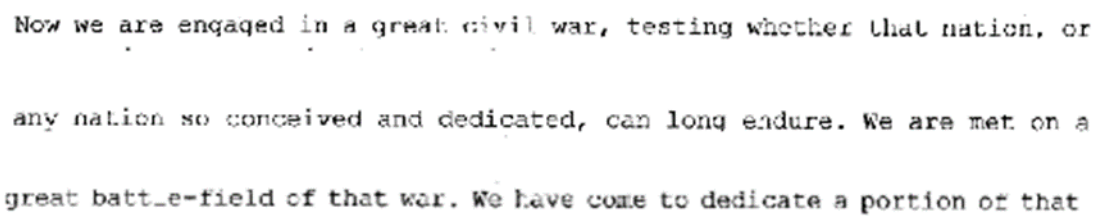

Figure. 4 Output of horizontal projection profile

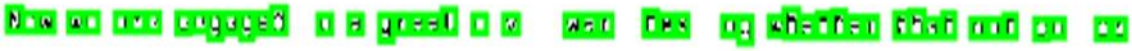

Figure.5 Separate Characters Bounded by a rectangle

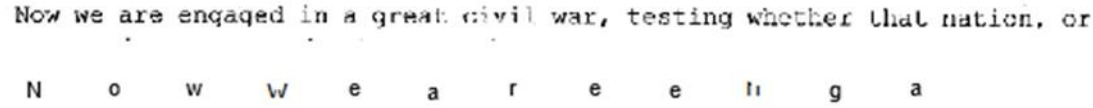

Figure. 6 Isolated Characters

A. Segmentation of Individual lines using Horizontal Projection Profile

In this we formulate a histogram to know where lays the gap between the lines based on the frequency of pixels occurring in the histogram and then set a threshold value to segregate separate lines from the original image. The output of horizontal projection profile is shown in Fig. 4.

B. Segmentation of Individual Characters using Vertical Projection Profile

In this method we formulate a histogram for the output images of horizontal projection profile in order to find the gap between the words and characters so that a threshold value is set based on which individual characters are segmented. The output of vertical projection profile is shown in Fig. 5. In Fig. 5 the output of horizontal projection profile is taken and individual characters are found and bounding boxes are drawn around each isolated characters using matlab inbuilt function.

Step 4 – From the output of Fig. 5 we extract individual characters by displaying each bounding box so that each character is displayed separately.

From the isolated characters from Fig 6 we apply chain coding to find out the exact boundary of the character using chain code. This is done to get the exact boundary so that we can then process it and get exact results out of it. The output of chain coding is then used in the next step to calculate the mean width of the character.

Step 5 – Mean based Thresholding method is used to identify the damaged characters. In this method we first take the width of full characters. The width of full characters is calculated from Eq. (1). From the width of full characters calculated a mean threshold value using Eq. (2) is found.

$$A = B - C \quad (1)$$

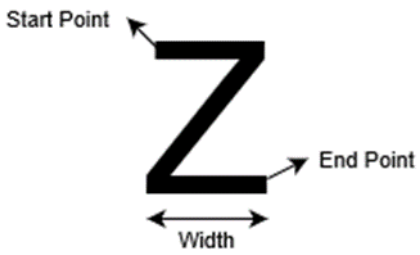

Figure. 7 Normal character width

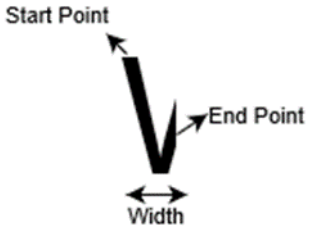

Figure. 8 Broken character Width

In which A stands for width, B stands for Endpoint of character and C stands for start point of character.

$$Mean\ Threshold = \frac{\sum A}{n} \quad (2)$$

Where A stands for sum of width of all the full characters and n stands for number of full characters.

Fig. 7 shows the width of full character and Fig. shows the width of broken character. From this it is known that normally the width of broken character will be much lesser than the width of the full character. By having all these values we check each and every character isolated from Fig. 6 and classify them as broken character or full character by using Eq. (3).

Step 6 – Once the broken characters are identified they are merged using the space information present near the character. If a character is broken vertically then it would be having two or more broken parts of it. These broken parts might be having a smaller space compared to the other normal characters which is shown in Fig. 9.

$$Character = \begin{cases} Full\ character; \\ if\ character\ width(A) \geq 95\% \\ of\ mean\ threshold, \\ Broken\ character; \\ if\ character\ width(A) < 95\% \\ of\ mean\ threshold \end{cases} \quad (3)$$

If a character is broken, then the first part of the character would be having a normal spacing from its previous character. But the next broken piece will be having a less space compared to other spacing which indicates that it is the broken piece of the previous character. The space between the characters is computed using Eq. (4).

$$Space = NS - PE \quad (4)$$

Where start point of next character is NS and End point of previous character is PE.

$$Space\ Before\ character = DCS - PCE \quad (5)$$

The space before damaged character is found by using Eq. (5), where DCS is damaged character start point value and PCE is previous character end point value.

$$Space\ After\ character = NCS - DCE \quad (6)$$

The space after damaged character is found by using Eq. (6), where NCS is next character start point value and DCE is damaged character end point value.

$$Threshold\ space = \frac{Minimum\ space}{Maximum\ space} \quad (7)$$

By using the above equations space between each character is found. From the spaces calculated minimum and maximum space value is taken and a threshold value is found using Eq. (7).

When space before character is less than the threshold space then it is combined with the previous character which is shown in Eq. (8).

$$Space\ Before\ Character \leq Threshold\ space \quad (8)$$

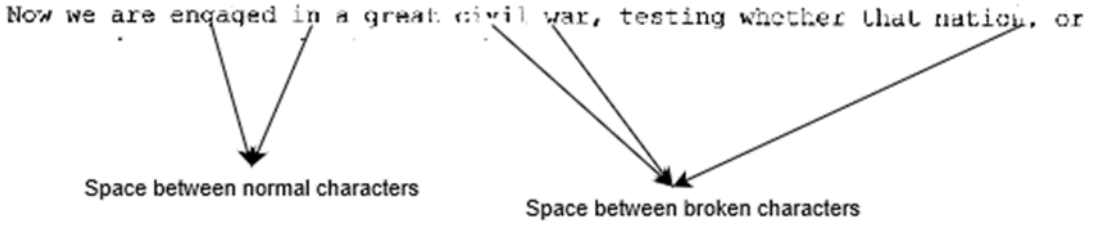

Figure. 9 Space between broken and normal characters

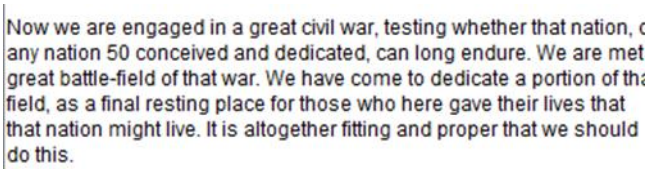

Figure. 10 Enhanced image

When space after character is less than the threshold space then it is combined with the next character, which is shown in Eq. (9).

$$Space\ After\ character \leq Threshold\ space \quad (9)$$

Based on this information the two broken pieces are merged together to form a whole character that makes the document clear. Fig. 10 gives the result of enhanced document after applying the proposed method of work.

5. Results and Discussion

The proposed approach is tested on various types of damaged English documents which were taken as input by scanning the images. The proposed approach attains accuracy of 92.88% with the damaged English characters with skew free documents. The following Table 1 shows the experimental result achieved in processing the damaged documents.

In the below table precision is calculated using the formula,

$$Precision = \frac{Total\ no\ of\ correct\ characters}{No.of\ correct\ characters} \times 100 \quad (10)$$

Recall is calculated using the formula,

$$Recall = \frac{Correctly\ merged\ characters}{Broken\ characters} \times 100 \quad (11)$$

In the above table time complexity is calculated based on the time taken to scan the entire document and the time taken to merge the broken characters. So the time precision is high for documents with more number of characters and also more number of broken characters.

The accuracy of the proposed method is found to be 92.88% which is compared with various approaches to rebuild or merge the broken characters which is shown in Fig. 12. Fig. 12 is obtained by plotting the accuracy of various methods with the proposed method to show the efficiency of the proposed method. When compared with other methods, proposed method outperforms well in terms of accuracy.

Fig. 11 shows the comparison of all the document images precision, recall and time complexity in which time complexity is plotted against secondary axes and precision and recall are plotted against primary axes.

In Fig. 12, MLFNC stands for Multi-Layer Feed forward Neural Network that was used in one paper and has achieved an efficiency of 68.33% and Mean Based Thresholding method achieves an efficiency of 79.93% and Polynomial Curve fitting method has achieved an efficiency of 89%. However the proposed method Mean Based Thresholding combined with chain coding has achieved an efficiency of 92.88%.

Table 1. Experimental result

Input Image	No. of characters	Broken characters	Correctly merged characters	Total no of correct characters	Precision in %	Recall in %	Time Complexity
Image 1	209	26	23	206	98.56	88.46	2 sec
Image 2	152	13	12	151	99.34	92.30	1 sec
Image 3	350	20	19	349	99.71	95.00	2 sec
Image 4	766	40	33	759	99.08	82.50	5 sec
Image 5	653	45	43	651	99.69	95.55	5 sec
Image 6	796	56	50	790	99.24	89.28	6 sec
Image 7	589	36	35	588	99.83	97.22	4 sec
Image 8	1032	190	180	1022	99.03	94.73	9 sec
Image 9	288	19	18	287	99.65	94.73	1.05 sec
Image 10	90	5	4	89	99.88	80.00	0.5 sec

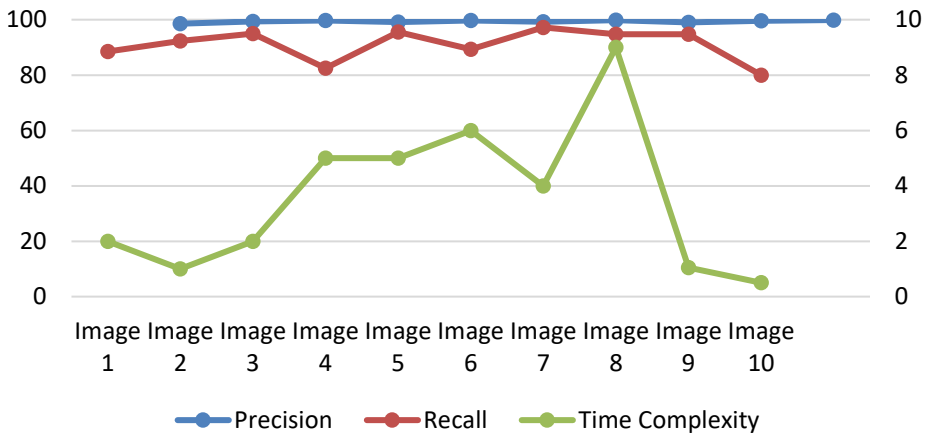

Figure. 11 Comparison of variations in precision, recall and time complexity

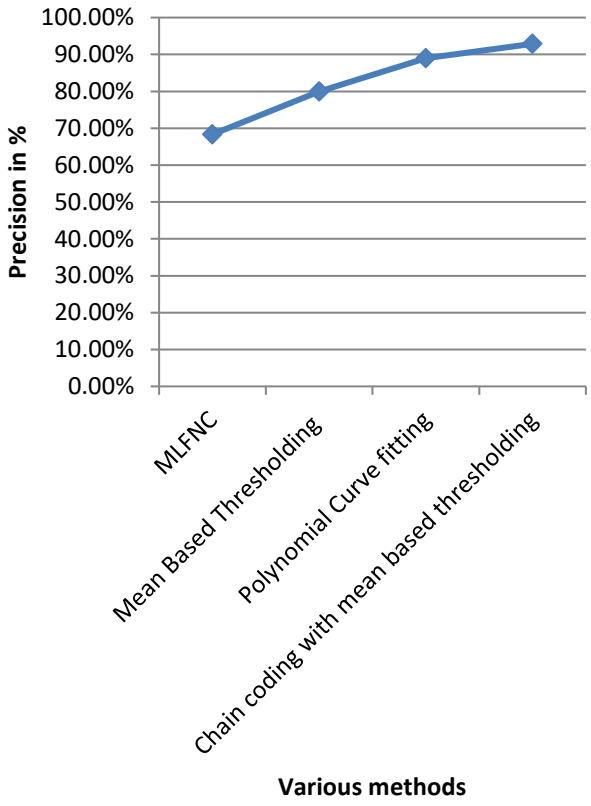

Figure. 12 Proposed model compared with other models

5. Conclusion

The approach proposed in this paper is tested in various types of scanned English documents with various fonts. The approach works well even for graphic images because of the effective preprocessing done on it. Though it works well for graphic images it does not work well for skewed images. The use of chain coding gave a better recognition result since the boundaries were correctly detected for each character. This increased the efficiency of recognition rates in the OCR machine and also added an advantage of digitizing the old documents very easily. This work can be further extended to recognize the broken characters in skewed documents also by using a very effective recognition system. Recognizing characters in a skewed documents is also present in some of the research papers that could be incorporated into this for an effective results.

References

[1] S. Bag and G. Harit, "A survey on optical character recognition for Bangla and Devanagari scripts", *Sadhana*, Vol.38, No.1, pp.133-168, 2013.

[2] S. Goyal and A.K. Bathla, "Method for Line Segmentation in Handwritten Documents with Touching and Broken Parts in Devanagari Script", *International Journal of Computer Applications*, Vol.0975–8887, 2014.

[3] Ø.D. Trier, A.K. Jain, and T. Taxt, "Feature extraction methods for character recognition-a survey", *Pattern recognition*, Vol.29, No.4, pp. 641-662, 1996.

[4] S. Kubatur, M. Sid-Ahmed, and M. Ahmadi, "A neural network approach to online Devanagari handwritten character recognition", In: *Proc. Of International Conf. On High Performance Computing and Simulation (HPCS)*, pp. 209-214, 2012.

[5] A.P. Whichello and H. Yan, "Linking broken character borders with variable sized masks to improve recognition", *Pattern Recognition*, Vol.29, No.8, pp.1429-35, 1996.

[6] M. Droettboom, "Correcting broken characters in the recognition of historical printed documents", In: *Proc. of Joint Conf. on Digital Libraries*, pp.364-366, 2003.

[7] J.R. Shah, and T.V. Ratanpara, "A Mean-Based Thresholding Approach for Broken Character Segmentation from Printed Gujarati Documents", In: *Proc. of the Second International Conf. on Computer and Communication Technologies*, Springer, New Delhi, pp. 487-496, 2016.

[8] K. Kaur and A.K. Bathla, "Segmentation of Degraded Text using Dynamic Profile Projection in Handwritten Gurmukhi Script", *International Journal of Engineering and Computer Science*, Vol. 5, No. 11, 2016.

[9] N. Sandhya and R. Krishnan, "Broken kannada character recognition—A neural network based approach", In: *Proc. Of International Conf. On Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp. 2047-2050, 2016.

[10] N. Sandhya, R. Krishnan, and D.R. RameshBabu, "A novel local enhancement technique for rebuilding Broken characters in a degraded Kannada script", In: *Proc: of Conf. on Advance Computing*, pp. 176-179, 2015.

[11] P. Mangla and H. Kaur, "An end detection algorithm for segmentation of broken and touching characters in handwritten Gurumukhi word", In: *Proc. of the 3rd International Conf. On Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, pp. 1-4, 2014.

[12] M. Yetirajam, M.R. Nayak, and S. Chattopadhyay, "Recognition and classification of broken characters using feed forward neural network to enhance an OCR solution", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, Vol.1, No.8, 2012.

[13] C. Sumetphong and S. Tangwongsan, "Effectively recognizing broken characters in Historical documents", In: *Proc. of International Conf. on Computer Science and Automation Engineering (CSAE)*, Vol.3, pp. 104-108, 2012.

[14] N. Premchaiswadi, W.P.U. Pachiyanukul, and S. Narita, "Broken characters identification for Thai character recognition systems", *WSEAS Transactions on Computers*, Vol.2, No.2, pp.430-434, 2003.

[15] S. Singh, A. Aggarwal, and R. Dhir, "Use of Gabor Filters for recognition of Handwritten Gurmukhi character", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol.2, No.5, 2012.

[16] T. Saba, A. Rehman, A. Altameem, and M. Uddin, "Annotated comparisons of proposed preprocessing techniques for script recognition", *Neural Computing and Applications*, Vol.25, No.6, pp.1337-1347, 2014.

[17] L. Yang and H.C.W. Zhang, "A License Plate Recognition Method for Community Monitor Based on Hausdorff Distance", *International Journal of Intelligent Engineering and Systems*, Vol.6, No.3, pp. 10-16, 2013.

[18] M.T. Parvez and S.A. Mahmoud, "Arabic handwriting recognition using structural and syntactic pattern attributes", *Pattern Recognition*, Vol.46, No.1, pp.141-154, 2013.

[19] A.S. Vaidya and B.R. Bombade, "A novel approach of handwritten character recognition using positional feature extraction", *International Journal of Computer Science and Mobile Computing*, Vol.2, No.6, pp.179-186, 2013.

[20] M. Vaidya, Y.V. Joshi, and M. Bhalerao, "Marathi numeral identification system in Devanagari script using discrete cosine transform", *International Journal of Intelligent Engineering and Systems*, Vol.10, No.6, pp.78-86, 2017.

[21] T. Saba, A. Rehman, and M. Elarbi-Boudihir, "Methods and strategies on off-line cursive touched characters segmentation: a directional review", *Artificial Intelligence Review*, Vol.42, No.4, pp.1047-1066, 2014.