JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.7, Special Issue.1, Jan 2019 E-ISSN: 2347-2693

Degraded Bangla Character Recognition by *k*- NN Classifier

Jayati Mukherjee[1*], S. K. Parui[2], Utpal Roy[3]

[1]Department of Computer and System sciences, Visva-Bharati, Santiniketan, India
[2]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India

[*]*Corresponding Author: jayati.uit93@gmail.com, Tel.: +91-9474481001*

Available online at: www.ijcseonline.org

***Abstract*—** Digitization of Bangla degraded document by Optical Character Recognition is a research activities now a days. Some historical documents particularly of 60s and 70s are degrading day by day due to lack of preservation. Those need to be retrieved. In this article, we present our recent study on recognition of degraded printed document images of Bangla, the 7th most popular language in the world. In the proposed approach the input will be low quality degraded images and the output is the recognized characters. In the first step some preprocessing are done on the document image to improve the quality of the scanned image. The proposed approach is an analytic approach. The segmentation is carried out line by line, word by word and finally character by character. The database used is the ISIDDI database. The total number of historical pages in TIF and JPG formats are 535, containing different fonts, sizes, formats and most importantly different levels of degradations. After segmentation we have manually identified 320 classes of such segmented symbols and divided the whole character dataset into test set (30%) and training set (70%). From the training set of 320 classes we have computed the Histogram of gradient feature or HOG feature on the samples. By applying the *K*-means clustering algorithm clusters for 320 classes has been generated and labeled according to the classes. For a character of test set again the HOG is computed and by applying *k*-nearest neighbour algorithm with the 320 classes the character is assigned to a character class with the minimum distance. The classification accuracy obtained on the test set is encouraging. We have achieved 82. 80% character or symbol level accuracy on 320 classes from the confusion matrix.

***Keywords*—** Degraded document recognition, Bangla document analysis, *K*-Means, *k*-nearest neighbour.

I. INTRODUCTION

Optical Character Recognition (OCR) is a process of recognition of printed or written text characters by a computer. This involves photo scanning of the document, analysis of the scanned-in image and then translation of the character image into character codes, such as ASCII, commonly used in data processing. Rigorous research on OCR is going on for few decades. In Indian languages like Oriya [1], Tamil [2], Bangla [3] [4], Gujrati [5], Malayalum [6], Telegu [7] the OCR systems on handwritten as well as printed documents, developed. Most of the concentration of research was upon good quality document images. There is no sufficient number of works on degraded documents specially for Indian languages like Bangla. Bangla is one of the most popular languages in the world. It is also official language of Bangladesh and some of the states of India like West Bengal, Tripura, part of Assam and part of Jharkhand. Bangla is the second most popular language in Indian subcontinent and Seventh most popular language in the world. In Bangla, the total number of characters is large but there is no upper and lower case differentiation as in English. Bangla is quite complex in nature compared to English. Here are some characteristics of Bangla script:

a) Bangla is written from left to right.
b) Bangla characters are connected by a headline called matra. Some characters are isolated in the word.
c) In Bangla, there are vowel modifiers and consonant modifiers. A vowel modifier occurs with a consonant to the left or right, or at the top or the bottom.
d) A consonant following or preceding another consonant changes shape which is a consonant modifier.
e) Compound characters are formed when two or more basic characters are combined together. In Bangla there are about 270 compound characters.

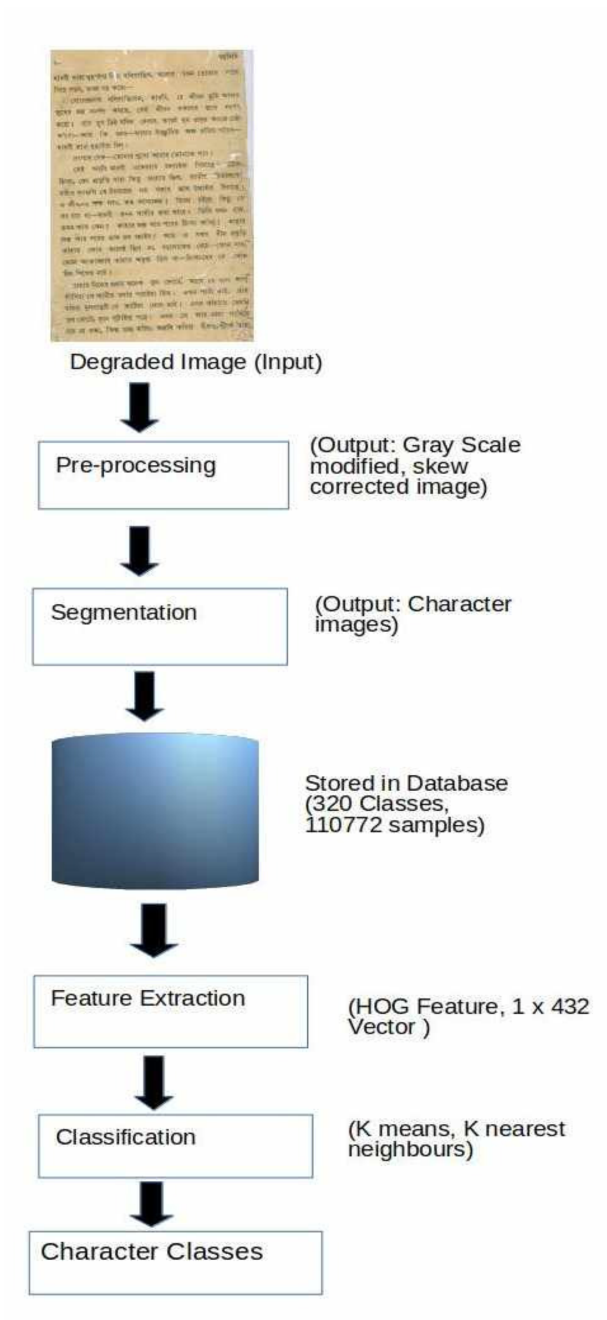

Degraded Image (Input)
Pre-processing
(Output: Gray Scale modified, skew corrected image)
Segmentation
(Output: Character images)
Stored in Database (320 Classes, 110772 samples)
Feature Extraction
(HOG Feature, 1 x 432 Vector)
Classification
(K means, K nearest neighbours)
Character Classes

Figure 1. Work Flow

Along with complex nature of Bangla another challenge is degradation. Degradation makes both the segmentation and recognition task difficult. There are different types of degradations we have faced in the Bangla document images like holes, stains, wrinkles, ink bleed. Physical adverse situation and natural calamities effect on the old books and degradation of different types can be seen on the books and documents [8] [9]. The degradation faced are as follows:

a) Change of colour: Yellowish background.

b) Noises, Stains, ink bleed, all these may be present in a degraded document.

c) Impression of the opposite side or bleed through is also a problem for historical documents.

d) Holes : Small holes may be present in a degraded document.

e) Broken or touching characters: Due to long lifetime part of characters may be faded or may be touched introducing broken or touching characters.

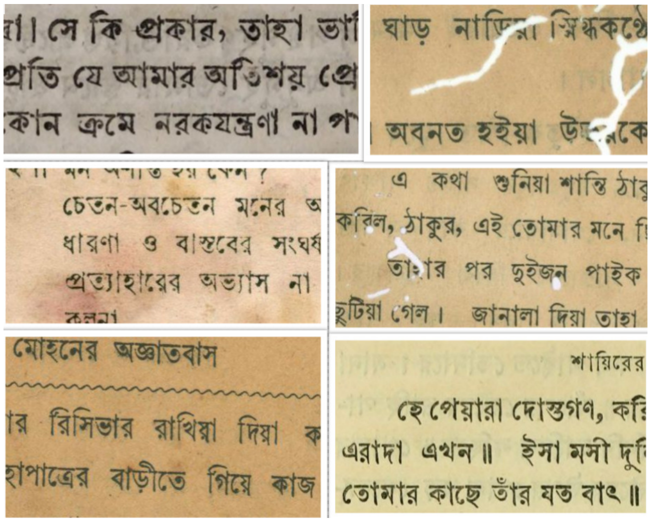
f) Skew: Due to misalignment during scanning of the document or the document may be already skewed.

The whole process of OCR is given in Figure 1. Pre-processing of the document is done to achieve an improved performance at the segmentation and recognition levels. OCR can be holistic or analytic. Segmentation is done on page images into lines, words and finally into pseudo-characters or individual symbols. From these pseudo-characters features are extracted and send to a classifier for recognition. For Bangla scripts it is quite difficult to segment and recognize the word images into characters or symbols due to the complex nature of Bangla language. Related surveys can be found in [10]. The task becomes more challenging when the document image is degraded due to unclear background, stains, ink bleed, holes, skew, broken characters, touching characters. For degraded documents some pre-processing are done to make the document legible to some extent. In our system we have pre-processed the document image. Segmented the page into lines and words and finally into characters. For recognition of character symbols we have applied *K* means clustering algorithm along with *k* nearest neighbour algorithm. From the ISIDDI database [11] we have manually identified 320 classes and we have proposed the scheme of recognition as a 320 class problem. Among 320 class 50 (39 vowels and 11 consonant) are considered for Bangla basic characters and rest 270 are considered as compound characters and punctuation marks.

The article organized as follows, section 2 explains the database we have worked with. Section 3 discusses the pre-processing techniques applied. Section 4 describes the recognition . Finally the results and discussions are in section 5.

II. Database

ISIDDI database [11]: The total number of historical pages, scanned using a flatbed scanner at 300 dpi and stored as color images in both uncompressed TIF and JPG formats, is 535 containing different fonts, sizes, formats and most importantly different levels of degradations. Part of pages images are given in Figure 2. Some of the printed degraded images are downloaded from the the Public Library of India (https://archive.org/details/digitallibraryindia). Some of them are extremely degraded and some have low level of degradations. The numbers of word and character classes appearing in this entire database are respectively 26,663 and 320.



া। সে কি প্রকার, তাহা ভা
প্রতি যে আমার অতিশয় প্রে
কোন ক্রমে নরকযন্ত্রণা না পা

ঘাড় নাড়িয়া স্নিগ্ধকণ্ঠে
অবনত হইয়া উ

চেতন-অবচেতন মনের অ
ধারণা ও বাস্তবের সংঘর্ষ
প্রত্যাহারের অভ্যাস না
কল্পনা

এ কথা শুনিয়া শান্তি ঠাকু
করিল, ঠাকুর, এই তোমার মনে
তাহার পর দুইজন পাইক
ছুটিয়া গেল। জানালা দিয়া তাহা

মোহনের অজ্ঞাতবাস
ার রিসিভার রাখিয়া দিয়া ক
হাপাত্রের বাড়ীতে গিয়ে কাজ

শায়িরের
হে পেয়ারা দোস্তগণ, কি
এরাদা এখন ॥ ইসা মসী দু
তোমার কাছে তাঁর যত বাৎ ॥

Figure 2. Raw Data

III. PRE-PROCESSING OF THE DEGRADED DOCUMENT

Pre-Processing of a document is an unavoidable step to be followed in case of degraded document. The pre-processing of the document creates an anomaly free and noise free document. A degraded document may has different types of degradation due to its long life time and adverse environmental conditions. So some pre-processing steps must be done to reduce the anomalies. The pre-processing phase consist of different stages. Mainly through accurate image thresholding, noise removal, skew detection/correction techniques, correcting broken or touching characters techniques the degraded document is pre-processed and prepared for segmentation and recognition.

Image Thresholding: The first step of pre-processing is thresholding. First it is changed to gray scale image, then we have applied sauvola's global thresholding method [12] to get a noise free binary image. Applying sauvola's global thresholding method we have computed a threshold value *T* on the scanned input image. The computed threshold value depends on the pixel values P_1, P_2, P_3 , …, P_i , ... , P_n of the image. The pixel values greater than the computed threshold is assigned 1 and the other pixel values lesser than the threshold is left as it is. So it has given a output image which contains a white background and a grey level fore ground. The character parts are gray level images.

Skew Correction: The perfect horizontal alignment of the image is important for segmentation and recognition of data. After scanning the document image skew may be available in the image. The skew correction method applied is based on the Hough transform. In [13] they have proposed a skew detection method on document image using Hough transformation.

Broken character correction: As we are working on historical documents, after being binarized it may contain broken pieces of the characters. The broken characters may affect the recognition process drastically and drop the recognition accuracy. So, the presence of broken characters may be a challenge in character recognition. The gray scale image of the historical document generally has a large amount of blurring. In addition, the pre-processing steps may introduce a large number of broken characters in the resulting binary image. We have tried to connect the broken characters to some extent using mathematical morphology [14].

Line Segmentation: For segmentation of words followed by segmentation of characters. Line segmentation is considered to be one of the important steps to be followed. As a conventional technique for text line segmentation global horizontal projection [15] of black pixels has been utilized.

Word segmentation: For segmenting the words we have applied mathematical morphology to trace the region of segmentation and extracted the words using connected components along with vertical projection method [15] [9].

Character Segmentation: For segmentation of characters only the headline of the words is removed. Using the horizontal projection method [15] the headline region is traced and removed. By removing the headline we have got some isolated character symbols. Those symbols are feed into classifier.

RecognitionIn recognition phase HOG fis used as the feature extractor from the character images. After feature extraction by applying the *K*-means clustering algorithm, clusters for 320 classes has been generated and labeled according to the classes. For a character of test set again the HOG is computed and by applying *k*-nearest neighbour algorithm with the 320 classes the character is assigned to a character class. The working diagram for recognition is given in Figure 3.

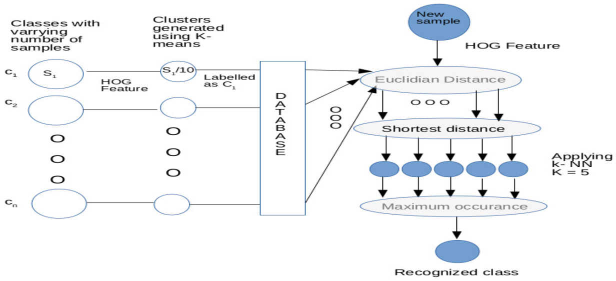


Figure 3. Working diagram for recognition

A. Database for recognition

We have created a database of 110772 samples with 320 classes from the ISIDDI database. We have manually classified them into 320 classes. The number of samples is not equal in each class. The number of samples in each class varies from 50 to 3000. For example, the class ক has huge number of samples due to frequent occurrence of ক in the Bangla documents and the occurrence of compound characters, like ঋ, are much less. The total samples are divided into two sets training set and test set dividing the samples into 77540 (70%), 33232 (30%) samples respectively. Some of the samples for training are given in Figure 4.

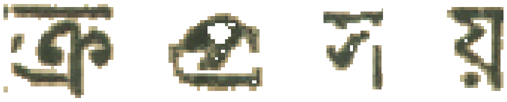


Figure 4. Samples of training set

B. HOG descriptor

In our work we have applied the HOG feature for extracting the Feature vector for recognition of Bangla character symbols. HOG feature gives an discriminatory result for each object. From the samples of the training set the feature vectors are extracted. The main idea behind the HOG descriptors is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions [16]. The implementation of these descriptors can be achieved by dividing the image into small rectangular regions, called cells, and for each cell, computing a histogram of gradient directions or edge orientations for the pixels within the cell. The combination of these histograms then represents the descriptor.

The horizontal and vertical gradient of the image is computed. Let the image be $I(x,y)$ where x, y denotes the position and $G_x (x,y)$ stands for the horizontal component and $G_y(x,y)$ denotes the vertical component. The horizontal and vertical components are defined as follows,

$$G_x(x,y)=I(x+1,y)-I(x,y) \quad (1)$$

$$\mathrm{G_y(x,y)=I(x,y+1)-I(x,y)} \quad (2)$$

The magnitude $M(x,y)$ and the value $V(x,y)$ of the gradient of the pixel (x,y) is computed as follows:

$$M(x,y)=(G_x^{\,2} + G_y^{\,2})^{1/2} \quad (3)$$

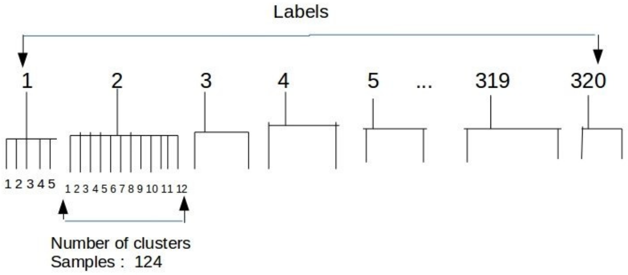
$$V(x,y)=(tan)^{-1}(G_x /G_y) \quad (4)$$

The computed value of magnitude and directions of the gradients determines which pixel will fall in which category. The categories are called Orientation bin(T), and for unsigned gradient evenly spaced between 0 to 180 degree and for signed gradient 0 to 360 degree. In our work the value of Orientation bin is 9. The pixel is assigned to the closest two bins to avoid noise which is determined by the linear interpolation or angular domain. The image is divided into sub-regions by non-overlapping window which moves from left to right. The magnitude are accumulated into histogram for each small sub region. At the end the histogram are combined to get the feature vector.

In our work the character size is different for each character so we have divided the characters into 4 * 5 cells. For the characters which are not divisible by 5 and 4 width wise and height wise respectively, a little padding is done. So we get total 12 blocks and 48 cells. The length of the feature vector is (48 * 9)=432. This feature vector is an input in the k-means clustering process.

C. K-Means

K-means clustering algorithm helps to create *K* number of disjoint cluster based on the number of observations [17]. The *K*-means clustering algorithm is applied on each class individually in the training set of 320 classes totallling 77540 samples. Here the number of samples in each class is considered to determine the value of *K* for that class. Let us consider the classes are named as C_1 , C_2, ... , C_n containg samples S_1, S_2, ..., S_n respectively. For C_i class the value of *K* will be, $K=round(S_i/10)$.



Fgure 5. Cluster Generation

So, in case of our Bangla degraded documents there are 320 classes and each class has different number of samples. After extracting the HOG feature for each datam we get 432 dimensional feature vector for each input datam. So total 432 dimensional 77540 feature vectors are used as the input samples. These 432 values together represent a point location in the 432 dimensional space. Our next aim is to build the clusters. Our database is organized into 320 classes and the samples are divided into their corresponding classes. Clusters are created for each of the class. The number of clusters for each class depend on the number of data samples. The clusters are labelled according to classes. For example, let a class labelled as C_i has 50 samples then 5 clusters will be generated and each of them will be labelled as C_i.

D. Classification by k nearest neighbours method

The recognition of a character sample is done by *k* nearest neighbours [18] algorithm. When the recognition phase comes, the HOG feature of the character is fed as the input. Let us consider *P* is a character sample. For *P,* feature vector is generated using HOG. It is also a 432 dimentional vector. Euclidean distance of the feature vector of *P,* with all the clusters generated from the *K-* Means algorithm is found. The Euclidian distance is measured by:

$$\sqrt{\sum_{i=1}^{N}(X_i - Y_i)^2} \qquad (5)$$

The Euclidian distance is the shortest geometric distance between two points. We have taken *k*=5 as the minimum number of clusters generated for a class is 5. Five nearest neighbous are considered from the feature vector. The label of the class that is assigned to maximum number of these 5 points or clusters, is selected as the class for *P,* the character sample.

IV. Result and discussions

The proposed approach for recognition of character samples from a degraded Bangla document has been tested on the test set of 320 classes on 33232 samples with different fonts, styles and degradation levels of ISIDDI database. From the confusion matrix of the classifier the top 15 accuracy in character level recognition is given in Table 1.

Table 1. *Top 15 character level accuracy of k-NN classifier obtained from the proposed strategy*

Sl. No.	Symbol	No. of samples	Recognized samples	Percentage(%)
1	□□□	118	109	92.37
2	□	1459	1328	91.02
3	□□□	126	114	90.47
4	□□□□	258	232	89.92
5	□	1271	1139	89.61
6	□	2350	2076	88.26
7	□	5865	5168	88.11
8	□	1801	1575	87.45
9	□	284	248	87.32
10	□	663	577	87.02
11	□	878	762	86.78
12	□□□	67	58	86.56
13	□	1019	879	86.26
14	□	98	84	85.71
15	□	1203	1025	85.20

The total accuracy on the dataset obtained is 82.80% on 33232 samples based on the confusion matrix. In future studies, we shall extend the simulation further by including a suitable language model which would increase the overall accuracy of the system.

References

[1] BB Chaudhuri, U Pal and Mandar Mitra, "Automatic recognition of printed Oriya script", Sadhana, Vol. 27, Pp. 23–34, 2002.

[2] R Seethalakshmi, TR Sreeranjani, T Balachandar, Abnikant Singh, Markandey Singh, Ritwaj Ratan and Sarvesh Kumar, "Optical character recognition for printed Tamil text using Unicode", Journal of Zhejiang University-SCIENCE A, Vol.6,Pp. 1297–1305, 2005.

[3] BB Chaudhuri and U Pal, "A complete printed Bangla OCR system", Pattern Recognition, Vol. 31, Pp. 531–549, 1998.

[4] Ujjwal Bhattacharya, Malayappan Shridhar and Swapan K Parui, "On recognition of handwritten Bangla characters", Computer Vision, Graphics and Image Processing, Springer publisher, Pp. 817–828, 2006.

[5] Apurva A Desai, "Gujarati handwritten numeral optical character reorganization through neural network", Pattern Recognition, Vol. 43 Pp. 2582–2589, 2010.

[6] Binu P Chacko, VR Vimal Krishnan, G Raju and P Babu Anto, "Handwritten character recognition using wavelet energy and extreme learning machine", International Journal of Machine Learning and Cybernetics, Vol. 3,Pp. 149–161, 2012.

[7] C Vasantha Lakshmi and C Patvardhan, "An optical character recognition system for printed Telugu text", Pattern analysis and applications, Vol. 7, Pp. 190–204, 2014.

[8] Kapil Dev Dhingra, Sudip Sanyal, and Pramod Kumar Sharma, "A robust ocr for degraded documents", In Advances in Communication Systems and Electrical Engineering, Springer publisher, Pp. 497–509 , 2008.

[9] Laurence Likforman-Sulem, Abderrazak Zahour, and Bruno Taconet. "Text line segmentation of historical documents: a survey", International journal on document analysis and recognition,Vol. 9(2), Pp. 123–138, 2007.

[10] Tapan Kumar Bhowmik, Swapan Kumar Parui, Utpal Roy, and Lambert Schomaker, "Bangla handwritten character segmentation using structural features: A supervised and bootstrapping approach", ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 15(4), Pages. 29, 2016.

[11] Chandan Biswas, Partha Sarathi Mukherjee, Koyel Ghosh, Ujjwal Bhattacharya, and Swapan K. Parui, "A hybrid deep architecture for robust recognition of text lines of degraded printed documents", In 24th International Conference on Pattern Recognition, IEEE, 2018.

[12] Jaakko Sauvola and Matti Pietikäinen, "Adaptive document image binarization", Pattern Pecognition, Vol. 33(2), Pp. 225–236, 2000.

[13] Chandan Singh, Nitin Bhatia, and Amandeep Kaur, "Hough transform based fast skew detection and accurate skew correction methods", Pattern Recognition, Vol. 41(12), Pp. 3528– 3546, 2008.

[14] Ying Jie Liu and Fu Cheng You, "Application of mathematical morphology on touching or broken characters processing", In Advanced Materials Research, Vol. 171, Pp. 73–77, 2011.

[15] BB Chaudhuri and U Pal, “A complete printed bangla ocr system”, Pattern Recognition, Vol 31(5), Pp. 531–549, 1998.

[16] Mohamed Becha Kaaniche, Francois Bremond, “Tracking HoG Descriptors for Gesture Recognition”, Advanced Video and Signal Based Surveillance, 2009 AVSS'09, Sixth IEEE International Conference on, Pp. 140–145, 2009, IEEE.

[17] John A Hartigan and Manchek A Wong, “Algorithm as 136: A k-means clustering algorithm”, Journal of the Royal Statistical Society. Series C (Applied Statistics),Vol. 28(1), Pp. 100–108, 1979.

[18] Keinosuke Fukunaga and Patrenahalli M. Narendra, “A branch and bound algorithm for computing k-nearest neighbors”. IEEE transactions on computers, Vol. 100(7), Pp. 750–753, 1975.

Authors Profile

Jayati Mukherjee has done her Bachelors degree in Computer Science and Engineering from University Institute of Technology, Burdwan University in 2015. She has done her Masters degree from BIT Mesra on 2017. She has worked as a PLP in Indian statistical Institute, Kolkata. Currently she is pursuing her Ph. D. degree from Visva Bharati, Santiniketan.

S.K. Parui received his Master's degree in Statistics and his Ph.D. degree from the Indian Statistical Institute in 1975 and 1986, respectively. He is now a Professor in Computer Vision and Pattern Recognition Unit of the Indian Statistical Institute. His current research interests include image processing, handwriting recognition, statistical pattern recognition, neural networks and information retrieval. Prof. Parui has published nearly 150 papers in refereed journals, edited volumes and conference proceedings. He has supervised several Ph.D. students.

After the completion of his Ph.D. from Department of Mathematics Visva-Bharati he went to the LAVAL University, Quebec, Canada for the Post Doctoral work in 1994. He worked at Indian Association for the Cultivation of Science, Jadavpur, Kolkata 32 as CSIR Scientist Pool during 1996-1997. He Joint the Visva-Bharati at the end of 1997 as Asst. Prof in Computer Science to teach the MCA course. He worked as Visiting Scientist in Academia Sinica, Taipei Taiwan during 2001 and 2002. He worked as Professor in IT in Assam University Silchar, Assam during 2008-2009. Presently he is a Professor and Former Head of the Department, Department of Computer & System Sciences, Visva-Bharati. He has been guiding Ph.D. students since long time and many students have been awarded Ph.D. under his supervision.