Intelligent Recognition of Characters from Ancient Manuscripts-A Review

Miss. Ketki. R. Ingole[1], Dr. Pritish A. Tijare[2]

[1]*Sipna College of Engineering & Technology, India*
[2]*Sipna College of Engineering & Technology, India*

Abstract

Now a days, Intelligent Character Recognition used for various application such as bank, school-colleges,online business. Intelligent Character Recognition used in handwritten document recognition, which written in different styles and format. Ancient manuscripts one of the area where automatic character recognition is required. India has prosperous collection of manuscripts, available at museums, National library in degraded form. A major steps taken by Government, Institutes and some organization for preserving these manuscripts. The research work has been done mainly in image processing area for preserving manuscripts, and it get succeed it. These manuscripts were written in various scripting languages and with different writing styles. Mainly manuscripts were written in continuous manner without picking the pen, which gives the character a cursive appearance. Character Recognition systems worked on the normal handwritten documents, but faces difficulties to read a cursive characters. With the help of Artificial Intelligence system are able to recognize characters by feature extraction and classification technique. The major hurdles of character recognition from ancient manuscripts are degraded manuscripts,different writing styles, similar character form. Details regarding recognizing characters from ancient documents are reviewed in this paper.

Keywords

Intelligent Character Recognition, Ancient Manuscripts, Artificial Intelligence, Feature Extraction and Classification,Degraded Manuscripts

1. Introduction

Ancient manuscripts is an repository of cultural heritage. It cultivates the knowledge in a form of scripts and in form of stories of different civilization from different era.India is identified as land of versatility in the form of culture heritage, and this heritage cultivated through this these ancient manuscripts. Indian Ancient Manuscripts stores knowledge of astronomy,cosmology, arts, medicine, mathematics and science and it cultivates from century to century and it passes from generation to generation.

In India temples, museums are source of such a manuscripts. In old age Indian Emperors, authorities of temples, were take charge of such manuscripts and degraded manuscripts always been get destroyed after they had been copied. As time passes, these cycle of restoration was broken, resulting into degradation of manuscripts and ignorance of knowledge it contains.

Beside this, India has richest collection of manuscripts,scatters at various Museums, National Libraries,Universities and temples, written in various languages such as Sanskrit, Tamil, Marathi and Telugu written in various scripting languages. But most of the manuscripts are available in degraded form and others

ACI'22: Workshop on Advances in Computation Intelligence, its Concepts & Applications at ISIC 2022, May 17-19, Savannah, United States
EMAIL: mohodketki@gmail.com

at the edge of degradation.Now a days, government, many foundations and many institutes come forward, and taking efforts for preserving these treasure of knowledge.Such organization are working for preserving these cultural heritage in digital form either by scanning or through digital photograph. With the help of digital image processing, manuscripts are not only get preserved but it can restore in their original form.
Most of the manuscripts were written on leaf, paper, and metal sheet.As the preservation cycle was broken, these manuscripts are facing problem of degradation. Mostly manuscripts get degraded due to natural end of life of paper and leaf, fungus attack, cockroaches, and these are the main hurdles while converting it into digital format. Now a days, advanced photography and scanning equipment are available, still the digital form of manuscripts are unreadable form. As the text in manuscripts are not clear due to ink seepage, dirt, wholes and cracks, while scanning manuscripts demanded to be flat as they are deteriorated , most of the time during image acquisition the light source is uneven, which leads to difficulties while extracting characters from digitized manuscripts.
With the help of image enhancement technique helps to enhance the digital form of manuscripts which leads to retrieval of text from such degraded manuscripts.It reduces the hurdles of text extraction and increases the readability.But still adequate methods are unrevealed for producing quality results.
Manuscripts were written in various scripting language, such scripts were written with ink, and in cursive and continuously. While recognizing text from such manuscripts cursive writing, writing style, uneven alignment, text without punctuation marks, creates problems. Most of the manuscripts were written double sided, which are facing a problem of bleed through effects.
Character recognition is a vast area of research and mostly work done on it, still ancient scripting language character recognition requires more focus. As there are few people are available for recognizing such ancient scripts and retrieving knowledge from it. As many organization and government working on it, but due to presence of overlapping lines, different writing styles,and similar shapes of text increases the complexity of recognizing characters.

2. Need To Explore The New Approach

Most of the manuscripts are unreadable due to ink seepage and background impression called as bleed-through effect. Due to bleed-through effect difficulties arises while separating the text from background manuscript image.The discoloration of manuscripts leads to produce the same color and hence texts are not clear to read.
To improve the readability of manuscripts various Image processing techniques are available, but still such methods are not adequate to produce a quality result for some documents. Most of the work is done on removing noise from the images and character recognition. Hence there is a scope to improve the manuscript image quality and make it available to researchers.
At the same time different handwriting is one of the hurdles to co-relate the characters.Pattern recognition is used to reduce the ambiguity for identifying characters/words from ancient documents.

3. Literature Review

To improve the quality of manuscript images and character recognition researchers have discovered many methods, a short report is presented in this section.
Previously designed manuscript image enhancement algorithms aims are to retrieve the text from manuscripts with rough background. At first, (N.Ostu,1979) proposed the thresholding method on gray scale images by increasing the discriminant measures between the pixel. This method is time consuming due to inefficient formulation and with fixed threshold it becomes difficult to achieve consistent quality.
The research work dedicated for colored document images. DjVu, (Bottou, 1998) , in context of compression implements an algorithm efficiently worked for separating foreground-background.
Most of the manuscripts written on both sides, while restoring such a document back impression is an major hurdle. The proposed direct image matching and directional wavelets methods reduces background noise and bleed through effect (Q.Wang, 2003).

The proposed adaptive method helps to separate the text by using local information. The proposed for low quality color document images (Leydier, Emptoz,2004).
The proposed (Zhixin Shii and Venu Govindraju, 2005) an algorithm uses a local adaptive binarization algorithm for background light intensity normalization and enhancing images.
In another research paper, (Shi,Setlur,Govindraju, 2005), proposed image enhancement algorithm for color image of palm leaf manuscripts.Palm leaf manuscripts are available in various libraries. The degraded version of Palm leaf manuscripts are preserved by applying various chemicals, which reduces the readability of that document.The images of an ancient, degraded palm leaf enhanced by using enhancement algorithm.
Furthermore, Fuzzy logic method is used for ancient documents (J.M.Gil,2006). They investigated method for identifying distance between letters and character styles. They used Gabor filter for feature extraction and for feature classification techniques fuzzy logic is used. With the help of Gabor filter local information is extracted from different environment in an image and aspect ration calculated for each character.
The hybrid method, (Wafa Bousellaa, 2008), for image segmentation of historical Arabic manuscripts. This algorithm combines normalization and clustering algorithm for light intensity normalization and foreground and background separation.
Furthermore, Dynamic Bayesian Networks(DBNs) method used for recognising the characters(Sulem, 2008). DBNs method efficient for recognizing broken characters increases readability of degraded and unclear character improved.
Most of the historical documents suffering the problem of back impression if they written on both sides. To restore such document, (Jie Wang, 2011) proposed a theory of restoration with the help of non-rigid registration method. A method based on directional wavelet helps to reduce the back impression. The evaluation shows 85.2% precision and improve the appearances of document.
A Chain Code based method (Chandure and Inamdar,2017), worked on both Devnagari and Modi vowels characters, A Chain Code based method in combination with BPNN, KNN, and SVM.It shows good results for Devnagari vowels than Modi vowels.

4. Issue Identified in Literature Review

The below table1 shows the various issues which will be observe within the previously developed system for the enhancement and character recognition of ancient manuscripts.

Table 1
Issued Identified in previous developed systems for ancient manuscripts enhancement and character recognition.

Author Name and Reference	Technique used	Issues Identified
(N.Ostu,1979)	Threshold Technique	It is difficult to achieve consistent quality with fixed threshold.
(Bottou, 1998)	DjVu	Separates the foreground and background, but results not available for an ancient documents.
(Q.Wang, 2003)	Directional Wavelet	Reduces background noise and bleed through effect, still strong background impression reduces the accuracy.

(Leydier, Emptoz,2004)	K- Means Clustering	The results on ancient documents was good but computationally expensive.
(Zhixin Shii and Venu Govindraju, 2005)	Normalisation technique	If intensity of foreground text and background are same then difficulties arrived for extracting text.
(Shi,Setlur,Govindraju, 2005)	Normalisation technique	Generates binarised image with text degradation
(J.M.Gil,2006)	Fuzzy methods	Automatic parameter detection system with good heuristic function is required
(Wafa Bousellaa, 2008)	Normalisation Technique with K-means Clustering	Improvement is required in segmentation process.
(Sulem, 2008)	Bayesian Networks	Efficient to recognize broken characters but accurate parameter initialization is required.
(Jie Wang, 2011)	Feature Extraction Select control points with matching features	Post processing method is required to recover the broken foreground text
(Chandure and Inamdar,2017)	Chain code histogram	Large data set for accurate character recognition is required.

5. Proposed Methodology

5.1. Image(Color/Gray) Enhancement

Generally, historical documents facing two types of deficiencies. First, the original document is in deteriorated condition and second problem is uneven background while converting the document into digital form.
The enhancement technique helps to improve the image quality from the low contrast image. Most image enhancement technique is available to reduce the uneven background which perform well for some historical document and removes the hurdles for extracting the text.

5.2. Character / Pattern Recognition

After enhancement the image gets enhanced but still some texts are unrecognizable. Feature extraction for character recognition becomes difficult. That part can be made readable by using pattern recognition.

Feature and property extraction from input data is accomplished through the use of a neural network. Salient features that are regular to a given degree of shift and shape variations or distortions are automatically extracted by neural network.

6. Conclusion

In recent times,character recognition from ancient manuscripts has major concern, as manuscripts are repository of knowledge.Researchers in recognizing characters from ancient documents domain confronted challenge of analyzing a accurate characters from it. Enhancement techniques helps to resolve the primarily problem of uneven background intensity. Foreground and background normalization techniques separates the background from foreground text, which increases the readability. With the help of neural network text features extracted, for recognizing characters. As the low contrast document, skewed images, cursive writing, different writing styles leads for a further improvement of accuracy.

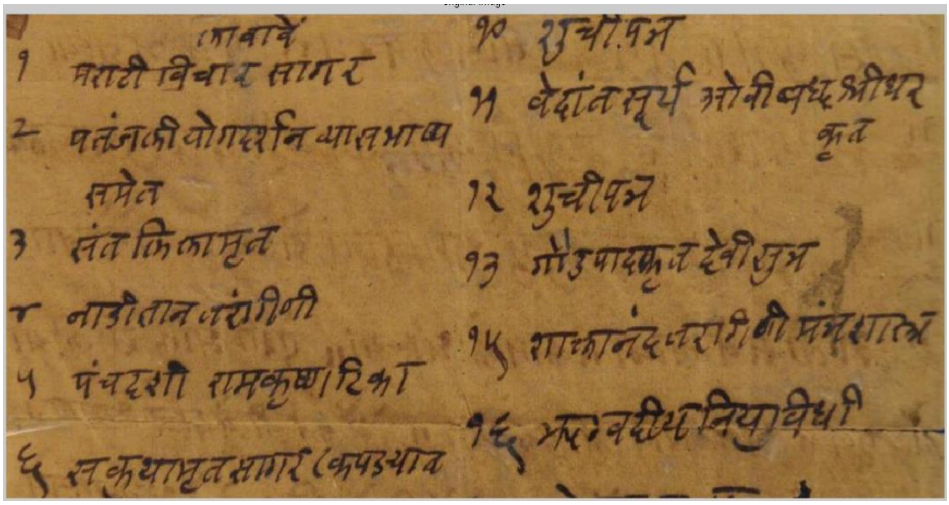

Figure 1: Ancient manuscripts

References

[1] N.Otsu, “A threshold selection method from gray level histogram",IEEE Transactions in Systems, Man, and Cybernetics, vol. 9, pp. 62, 1979.

[2] L.Bottou, P.Haffner, and P.G Howard, “High Quality Document Image Compression with DjVu”, Journal of Electronic Imaging, July 1998.

[3] Q. Wang, T. Xia, L. Li, and C. Tan, “Document image enhancement using directional wavelet," IEEE Conference, Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, June 2003.

[4] Y. Leydier, H. Emptoz and F.L. Bourgeois,“Serialized K-Means for Adaptive Color Image Segmentation-Application to Document Images and Others”, Workshop on Document Analysis Systems(6th International) , Italy, 2004.

[5] Zhixin Shi and Venu Govindaraju “Historical Handwritten Document Image Segmentation Using Background Light Intensity Normalization”, SPIE Document Recognition and Retrieval XII, San Jose, California, USA, January 2005.

[6] Venu Govindaraju, Zhixin Shi Srirangaraj and Setlur ,“Digital Image Enhancement using Normalization Techniques and their application to Palm Leaf Manuscripts”, CEDAR, Buralo,U.S.A.February, 2005 .

[7] J. M. C. Sousa, J. M. Gil, C.S. Ribeiro and J.R.C. Pintom, “Old Document Recognition Using Fuzzy Methods”, Intelligent Systems Technologies and Applications, Vol.1, 2006
[8] Wafa Boussellaa, “A Methodology for the Separation of Foreground/Background in Arabic Historical Manuscripts using Hybrid Methods”, Journal of Universal Computer Science, vol. 14,2008.
[9] M. Sigelle and L. L. Sulem, “Recognition of degraded characters using dynamic Bayesian networks”, Pattern Recognition, Vol. 41 Issue 10, 2008.
[10] Jie Wang; Chew Lim Tan; , "Non-rigid Registration and Restoration of Double-Sided Historical Manuscripts," Document Analysis and Recognition (ICDAR), 2011 International Conference, September 2011.
[11] V. Inamdar and S.L. Chandure, “Performance analysis of handwritten Devnagari and MODI Character Recognition system”, in Conference Computer Analysis of Secure Trends, 2017.