ISSN 2278-9723
DOI:10.31838/ijccts/09.01.05

**Review Article**

# Digitization, Preservation and Character Recognition in Ancient Documents Using Image Processing Techniques – A Review

**M.VAMSHI KRISHNA[1], K. JANAKI RAM[2]**
[2]Assistant Professor, Vignans Deemed to be University
Email: mvk.6518@gmail.com[1] Janakiram006@gmail.com[2]
Received: 18.07.21, Revised: 03.08.21, Accepted: 01.09.21

**ABSTRACT**
Historical documents are considered a significant source of national heritage and societal development. It is an essential feature of society and a reference to their culture, tradition, and civilization. An imagebased information management system for the restoration of cultural heritages has been developed. Precise records of the status of a cultural heritage are necessary to preserve or restore it. The precise records of the object have been available by photogrammetric technique, but there are too few expensive photogrammetric instruments and experienced photogrammetrists at a heritage site.

**Keywords**: Historical Documents, restoration, cultural heritage, photogrammetric instruments, photogrammetrists, OCR.

## 1. Introduction

A lot of precious old cultural heritages remain all over the world. These heritages are of great value for the human being in both history and art. Preserving the historical documents can be considered as preserving the culture of the heritage. The digital image database of historical documents is growing in the field of heritage studies. The work requires those images which are restored, enhanced, and stored reasonably to simplify access and disseminate.

Unfortunately, these historical documents confront the physical degradation caused by a combination of factors such as temperature levels, environmental conditions, and low-quality paper.

Archives and Libraries around the world store a plentiful of old and historically important documents and manuscripts. These contain a notable amount of heritage.

Many factors like improper handling, poor material quality used for their creation, ageing etc., make them suffer a high degree of degradation. Today as the world is moving towards digitization, we need to preserve their content for future generations.

Since the degraded historical document images are considered a combination of multilayer-information including the foreground (object) layer, background layer and the degraded layer. The image processing techniques can be applied to restore and enhance the quality of these degraded document images. The digital data produced requires 3 stages namely, 1) Pre-processing, 2) Enhancement and 3) Recognition. The main step in the whole process is binarization, which has the drawback of showing the influence on the quality of characters recognized. In addition to difficulties like faded ink, uneven illumination etc., there are also differences in patterns of handwritten and machine printed documents which are associated with binarization of old document images. Another challenging problem is word indexing. The purpose of indexing is to find the occurrences of a selected word within scanned images, without performing standard translation to text information.

The rest of the document is organized as follows. Section
2 describes the different approaches followed by different researchers to digitize and preserve the documents. Section 3 concludes the approaches and presents the future scope of the work.

## 2. Literature Survey

Although various technologies have been attempted to preserve or restore old cultural heritages, it is most important to record the status of the object precisely and preserve or restore histories of them accurately. These records are necessary to monitor the status of both damaged parts and restored parts of the target. A restoration researcher can make an appropriate preservation or restoration plan based on these records. However, precise, and accurate records of an old cultural heritage are not necessarily available. Precise records of the status of the object have been available by photogrammetric technique. But

this photogrammetric work is manual labor and requires a great deal of time and cost. Conventional photogrammetric instruments are too expensive and there is no experienced photogrammetrist working at many heritage sites. Therefore in [1], they developed an amateur system for recording the status of an old cultural heritage by digital cameras, assisting a restoration researcher to make appropriate preservation or restoration plans, and then managing restoration information such as date, position, treatment method , used chemicals and so on. The system was requested to be such as a restoration scientist without photogrammetric or image processing know-how can operate the system with shortperiod training. Furthermore, since information such as position and extent of damages has been managed usually on an analog map and/or analog inventory, this has made restoration research inconvenient to make an appropriate preservation or restoration plan.

### 2.1 Hardware Method

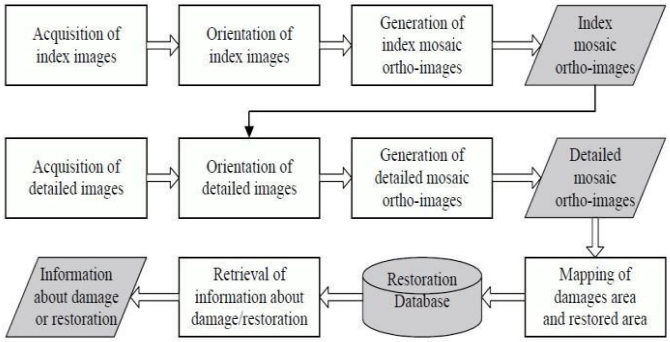


**Fig.1:Flow of standard processing**

#### 2.1.1 User Requirements

User requests to an information management system for restoration of cultural heritage.

1. A system should be operated easily by anamateur.
2. Hardware of a system should be compact & its cost should be low.
3. A system should be able to manage a great deal& diverse information.
4. Two-stage image acquisition

#### 2.1.2 Processing Flow

The processing flow of the system is like one of a conventional method by analog/analytical photogrammetry.

1. Image acquisition
2. Orientation of images
3. Generation of mosaic orthoimages
4. Mapping of the damaged area and restored area on a mosaic detailed image
5. Retrieval of information about damage

### 2.2 Software Methods

Feature extraction methods in [2] play an important role to improve the accuracy of the signature verification system. The pre- processing stage affects the accuracy and reduces computational time. The input image is first converted from RGB to Gray. Thinning is applied to the converted signature image. In the thinning process, the width of the signature is reduced which spontaneously removes the noise produced during the generation of the image.

In [3] presents the description of the method based on an adaptive multilayer-information binarization for restoration of degraded historical document images. The method has 5 stages. They are Noise removal using Wiener filter, Majority pixel analysis stage to extract foreground pixels from binary images, degradation of background layer by replacing foreground area stage, thresholding stage to transform gray level image into a binary image and the Vicinity analysis stage enhances the quality of a binary image by analysing and categorizing the pixels in the binary image. In [4] the image gradient has been widely used for edge detection and it can be used to detect the text stroke edges of the document images effectively that have a uniform document background. On the other hand, it often detects many non-stroke edges from the background of degraded document that often contains certain image variations due to noise, uneven lighting, bleed-through, etc. To extract only the stroke edges properly, the image gradient needs to be normalized to compensate the image variation within the document background.

In the proposed technique, first, an adaptive contrast map is constructed for an input image degraded

badly. Then the binarized contrast map is combined with an edge map, obtained from a canny edge detector to identify the pixels in the edges of the text stroke. By using local threshold the foreground text is further segmented which is based on the intensities of detected text stroke edge pixels within a local window.

To analysis the Filtering algorithms mentioned in [4], the below stated steps are followed: a) First, an uncorrupted document image is taken as input b) Next the document image is converted to RGB to gray image. c) Different noises are added to the document image artificially with 10% noise density. d) The filtering algorithms are applied for the reconstruction of document images.

f) To test the performance of the filters for varying noise density, Gaussian noise with different variance is applied on the image.

The steps involved in preprocessing using Neural Networks proposed in [8] are given below.

1. **Size normalization**: Bicubic interpolation is used for the standard sized image.
2. **Binarization**: it is the process of converting a grayscale image into a binary image by thresholding.
3. **Smoothing**: the erosion and dilation smooth the Boundaries of objects.
4. **Edge detection**: Morphological gradient operators are used in edge detection because they enhance the intensity of edges

of characters. The characters are always written in "print fashion", not connected, horizontal histogram profile (for line segmentation), and vertical histogram profile (for word segmentation).

A neural network method proposed in [6] is a powerful data modelling tool that can capture and represent complex input/output relationships. The motivation for the development of neural network technology stemmed from the desire to develop an artificial system that could perform "intelligent" tasks similar to those performed by the human brain. Neural networks resemble the human brain in the following two ways:

1. A neural network acquires knowledge through learning.
2. A neural network's knowledge is stored within interneuron connection strengths known as synaptic weights.

The square tracing algorithm a new approach explained in [7] is very simple this could be attributed to the fact that the algorithm was one of the first attempts to extract the contour of a binary pattern. Given a digital pattern i.e. a group of black pixels, on a background of white pixels i.e. a grid; locate a black pixel and declare it as your "start" pixel. (Locating a "start" pixel can be done in several ways; we'll start at the bottom left corner of the grid, scan each column of pixels from the bottom going upwards starting from the leftmost column and proceeding to the right- until we encounter a black pixel. We'll declare that pixel as our "start" pixel).

The first step proposed in [9] is to extract clusters of duplicate texts from the original collection of OCR'd documents. This would limit the applicability of our approach, and so instead of focus on detecting instances of local text reuse. Machine learning techniques like classification has been analysed in [10]. A supervised classifier is used to classify the text in the images. A Support Vector Machine is used to classify the image. Multiple symbols like alpha numerals, Hindi letters and special characters were combined to verify the classifier. Comparatively better results were obtained by using machine learning techniques.

In [11], Optical Character Recognition is used to segment the numbers present on a car license plate. The challenge is detecting the whole license plate and recognizing the characters present on the plate accurately. SVM classifier is used to classify the dataset. Histogram of the images are normalized for L1 multiclass SVM is trained. SVM takes the set of training features.

Optical Character Recognition (OCR) is difficult to perform in Indian languages [12]. A lot of research has been done in this regard but not in the context of the Indian language. Thirteen resource centres for Indian language technology solutions have been established at various educational institutes and research organizations. Relatively less work is available for the recognition of Indian languages. A possible reason could be the complexity of the shape of Indian scripts and some of the peculiarities in Indian scripts. A neural network-based script identification system is described in [13]. The system consists of a feature extractor and a neural network. The feature extractor produces the features employed by a probabilistic neural classifier. Feature extraction aims to describe the pattern with fewer features that helps in discriminating among pattern classes. The neural network discussed in this paper has 3 classifiers each to classify Engish, Hindi and Kannada languages. The classifier proposed in this paper solved the language identification problem.

In [14] OCR for the Telugu language has been proposed. The complexity of Telugu language is that it is more of curves than lines. So, letters in the script are treated as symbols to identify them. Direction features are stored from the set of training images. Then these features are compared with the features taken on the test images dataset. These features are compared using the k-nearest

neighbour classifier. Telugu letters are written in different fonts and sizes to classify them. Then these are classified using a KNN classifier. The overall recognition accuracy was approximately 92% for most of the images.

Another method for script recognition especilally in the Telugu language is proposed in [15]. In this, the general characteristics of the language are described. The main problem here with Telugu script compared with other languages is Telugu or any other Indian language has a larger set of characters. In contrast to cursive English script, characters are not connected in Telugu script. Another factor with the Telugu language is that some characters have similar shapes and can be differentiated only with some minor differences like some pen lifts or dots.

Telugu characters are visualised in terms of segments either in the form of a straight line or parts of circles. Based on the feature points, circular segments are characterized. These features are considered in the form of vectors which are known as feature vectors. Next is a character dictionary which has as the feature vectors of standard characters. Finally, the feature vector for the test sequence is compared with the dictionary entries. This way the test images and the dictionary images are matched.

### 3. Conclusion

As discussed in this paper there are different methods in digitizing ancient documents. While digitizing the documents, it is also important to recognize the language of the characters present in the document. In the character recognition system, very little research work is done on the concept of Indian languages. India being a land of multiple languages, there is a lot of scope in this area to work with. Pattern recognition in Indian languages is still in the early stages compared with English and other foreign languages. A lot of work is done in recognizing the printed scripts, but with handwritten scripts, it is still a long way to explore. Hence this area has a wide scope for research.

### References


1. Kenji Hongo, Ryuji Matsuoka, Seiju Fujiwara, Katsuhiko Masuda, and Shigeo Aoki, “Development of Image-Based Information System for Restoration of Cultural Heritage”, International Archives of Photogrammetry and Remote Sensing. Vol. XXXIII, Part B5. Amsterdam 2000.
2. Prof. Laxmikant Malphedwar, Mayur C. Waghere, Pratik V. Nimodiya, Nayan B. Mashalkar, “Survey on Offline Handwritten Signature”, IJETCS Vol 2, Issue 1, 2017.
3. G.Silpalatha, K.S.Raghavendra Reddy and B.Rajani Kumar Reddy, “Binarization of Document Image”, IJERA, Vol 6, Issue 5, May 2016.
4. Varada V M Abhinay, P. Suresh Babu, “A Novel Document Image Binarization For Optical Character Recognition”, IJCATR, Vol 3, Issue 9.
5. Reka Durai , Dr.V.Thiagarasu “A Study and Analysis on Image Processing Techniques for Historical Document Preservation”, IJIRCCE Vol. 2, Issue 7, July 2014.
6. Sameeksha Barve “Optical Character Recognition Using Artificial Neural Network”, IJARCET, VOl 1, Issue 6, June 2012.
7. Kandula Venkat Reddy, D.Rajeswara Rao, K. Rajesh, “Handwritten Character Detection by Using Fuzzy Logic
8. Techniques”, IJETAE, Vol 3, Issue 3, March 2013.
9. Ankit Sharma, Dipti R Chaudhary, “Character Recognition Using Neural Networks”, IJETT, Vol 4, Issue 4, April 2013.
10. Shaobin Xu, David Smith, “Retrieving and Combining Repeated Passages to Improve OCR”, IEEE 2017.
11. Vohra, Ujwal Singh, Shri Prakash Dwivedi, and H. L.
12. Mandoria. "Study and analysis of multilingual handwritten characters recognition using SVM classifier." Oriental J. Comput. Sci. Technol 9.2, 109-114, 2016.
13. Madhukar, R. B., Gupta, S., & Tiwari, P. (2015). Automatic car license plate recognition system using multiclass SVM and OCR. International Journal of Engineering Trends and Technology, 30 (7), 369-373.
14. Kumar, Anubhav & Sharma, Anuradha & Chawla, Monika & Prasad, T. (2009). Different Approaches in OCR of Indian Languages. 10.13140/RG.2.1.3243.5046.
15. Patil, S. B., & Subbareddy, N. V. (2002). Neural networkbased system for script identification in Indian documents. Sadhana, 27(1), 83-97.
16. C. V. Lakshmi and C. Patvardhan, "A multi-font OCR system for printed Telugu text," Language Engineering Conference, 2002. Proceedings, Hyderabad, India, 2002, pp. 7-17.
17. Rao, P. V. S., & Ajitha, T. M. (1995, August). Telugu script recognition-a feature-based approach. In Proceedings of 3rd International Conference on Document Analysis and Recognition (Vol. 1, pp. 323-326). IEEE.