A Systematic Approach for Various Phase Based Binarization and Segmentation of Ancient Document Images

Ritika Sood¹
Department of Electronics and Comm. Engg,
CGCCOE Landran, Punjab, INDIA¹ ²
soodritika6@gmail.com

Mr. Manvinder Sharma²
Department of Electronics and Comm. Engg,
CGCCOE Landran, Punjab, INDIA¹ ²

Abstract—At present era, globalization and technology improves the growth of any country. With the passage of time and new technology there is need to maintain and care of various historical monuments, culture and civilization and has to learn about these to all of the country people. Since they explain about the historic events so they play a vital role in determining the culture and civilization through which the nation outperforms a special place in all around the world. With the various rush in life many people are not familiar these historical and ancient documents. Today is the time of modernization but our culture and civilization belongs to ancestral. So, there is need to work over these areas to uplifts the culture and civilization. In this paper, the main concern is about phase based binarization of ancient document images. The dataset collected from different areas to emphasize the various historical events occur in Ancient India. Binarization is the process in which we enhance the quality and features of any images. The proposed model consists of three standard steps: 1) preprocessing; 2) main binarization; and 3) postprocessing. In the preprocessing and main binarization steps, the features used are mainly phase derived, while in the postprocessing step, specialized adaptive Gaussian and median filters are considered. The proposed work was quiet reliable and optimist in terms of PSNR and RMSE and found that the final overall output is quiet superior as compare to initial point.

Keywords— Binarization, Zoning algorithm, Root Mean Square Error, Peak Signal to Noise Ratio

I. INTRODUCTION

Ancient documents accumulate a significant amount of human heritage over time. However, many environmental factors, improper handling, and the poor quality of the materials used in their creation cause them to suffer a high degree of degradation of various types. Today, there is a strong move toward digitization of these manuscripts to preserve their content for future generations. The huge amount of digital data produced requires automatic processing, enhancement, and recognition.[2]The representation of ancient documents in digital databases has been considered in several works covering issues, such as color management, the accidental introduction of artifacts into images and the acknowledgement by the scholars of the quality of digitised images.[6] Ancient and historical documents, printed or handwritten, strongly differ from the documents mentioned above because layout formatting requirements were looser. Their physical structure is thus harder to extract. In addition, historical documents are of low quality, due to aging or faint typing. They include various disturbing elements such as holes, spots, writing from the verso appearing on the recto, ornamentation, or seals. Handwritten pages include narrow spaced lines with overlapping and touching components [7]. Historical records of daily activities provide intriguing insights into the life of our ancestors, useful for demography studies and genealogical research. Automatic processing of historical documents, however, has mostly been focused on single works of literature and less on social records, which tend to have a distinct layout, structure, and vocabulary. Such information is usually collected by expert demographers that devote a lot of time to manually transcribe them. To support research in automatic handwriting recognition for historical documents containing social records.

In recent years, automatic analysis of historical document images has received great interest. The need of automatic processing and archiving of large volumes of old documents and manuscripts, attracted many researchers. Preprocessing and post processing steps significantly improve the performance of binarization methods, especially in the case of severely-degraded historical documents. The proposed work mainly deals with step by step binarization and segmentation of ancient document images and improved the feature vectors using Zoning algorithm.
II. RELATED WORK

Most of the phase based binarization algorithms in the literature can be classified into region-based and connected component (CC)-based approaches. Region-based methods adopted a sliding window scheme, which is basically a brute force approach which requires a lot of local decisions. Therefore, the region-based methods have focused on an efficient binary classification (text versus non-text) of a small image patch. Phase base binarization is of fundamental importance in image understanding and content based retrieval. For instance the binarization and segmentation of images are quiet essential terms. Stability of such method includes robustness to noise and blur, because they accomplish features assembled throughout the region of interest. The second approach used is segmenting the entire image using the local parameters of an image (morphology, dilation, segmentation, thresholding, feature extraction).

Ancient and historical documents, printed or handwritten, strongly differ from the documents mentioned above because layout formatting requirements were looser. Their physical structure is thus harder to extract. In addition, historical documents are of low quality, due to aging or faint typing. They include various disturbing elements such as holes, spots, writing from the verso appearing on the recto, ornamentation, or seals. Handwritten pages include narrow spaced lines with overlapping and touching components. Characters and words have unusual and varying shapes, depending on the writer, the period and the place concerned. Many researchers have made research related to this but no technique is almost perfect and they found there should be need to improve the work in more areas at different instants and techniques.

III. DESIGN AND IMPLEMENTATION

The proposed research started with the objective of processing and refining image dataset that we are using a in the proposed framework and algorithm and in this process, following steps and processes evolved which lead to development of the research work.

Flow chart of proposed work:
To find the proposed objectives the proposed works mainly work upon Zoning algorithm.

**Zoning feature extraction algorithm:**

Zoning features have been proved one of the most efficient statistical features which provide high speed and low complexity word matching. They are calculated by the density of pixels or pattern characteristics in several zones that the pattern frame is divided. In the proposed work, an adaptive zoning technique is introduced. The main idea is that the zoning features are extracted after adjusting the horizontal boundaries of the zones with the use of Dynamic Time Warping (DTW). This adjustment is performed by coupling every zone of the query word to the corresponding zone of each candidate match-word and pixels of cells with the use of the corresponding warping matrix. This process absorbs the ambiguities between the query and the dataset images.

**Feature Extraction by Zoning Method:**

For feature extraction Image Centroid and Zone (ICZ) based Distance metric feature extraction system is used. Algorithm for that is as follows:

Input: Ancient Document image.

Output: Feature for Classification and Recognition.

Method Begins

Step 1: Compute the input image centroid.

Step 2: Divide the input image into n zones.

Step 3: Compute the distance between the image centroid to each pixel present in the zone.

Step 4: Repeat the step 3 for the entire pixel present in the zone.

Step 5: compute average distance between these points.

Step 6: Repeat this procedure sequentially for the entire zone.

Step 7: Finally n such features will be obtained for classification and recognition.

**Architecture of Zoning Feature Extraction:**

```
Input Image

Data Acquisision/ Preprocessing

Feature Extraction by Image Zoning

Verification on the basis of entire functions

Knowledge Base
```
There are various process and implementation made to design our proposed method:

**Collection of Dataset-Selecting highly relevant Image Instances to collect datasets:**
As the work was done on domain specific area, it was important that only relevant images instance datasets were picked. In this phase we collect dataset from different aspects and varieties of ancient and historical document images.

**Image Pre processing**
In order to conveniently compare to different contrast and different distinctive size of dataset images and consider the PC's velocity, the span of every one of these images ought to be restricted inside suitable pixels.

First, it's necessary to choose a proper format of input images and their appropriate dimensions. After that the actual procedure of pre-processing phase taken under process. Since, the proposed work is highly based over systematic approach for various phase based binarization and segmentation. For the sake of this, the various phases of pre-processing like conversion of color image into gray scale, thinning of image, binarization of image, edge detection, morphology and dilation of image would be separately for horizontal and vertical axis both.

**Removal of Noise- Post Processing Phase**
Since, noise in any image degrades the quality of image. So, if there will be noise then we have to remove it first. This phase will be done after the total procedure of pre processing.

**Segmentation Phase**
In this phase, we have to segment the characters or image in the image by bounding box method or by bounding the characters by canny edges detection method and after that finally we are in position to segment each and every part of text or image of the input taken.

**Feature Extraction:**
The objective of feature extraction is to capture the essential characteristics of the symbols and it is generally accepted that this is one of the most difficult problems of pattern recognition. The most straightforward way of describing a character is by the actual raster image. Another approach is to extract certain features that still characterize the symbols, but leaves out the unimportant attributes. The techniques for extraction of such features are often divided into three main groups, where the features are found from:

- The distribution of points.
- Transformations and series expansions.
- Structural analysis.

**IV. RESULTS AND DISCUSSION**

**Evaluation Parameters**

Two of the most important measures used to evaluate the research work are as follows.

1. RMSE (Root Mean Square Error)
2. PSNR (Peak Signal to Noise Ratio)
The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed.

\[
\text{RMSD} = \sqrt{\frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}}.
\]

The RMSE is the square root of the variance, known as the standard deviation. On the other hand, Mean Square error is only the variance for the estimator.

PSNR is generally used to analyze quality of image, sound and video files in dB (decibels). In other words, PSNR is to measure the quality of reconstructed images that have been compressed.

PSNR is most easily defined via the mean squared error (MSE). Given a noise-free \( m \times n \) monochrome image \( I \) and its noisy approximation \( K \), MSE is defined as:

\[
\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2
\]

\[
\text{PSNR} = 10 \cdot \log_{10} \left( \frac{\text{MAX}_I^2}{\text{MSE}} \right)
\]

\[
= 20 \cdot \log_{10} \left( \frac{\text{MAX}_I}{\sqrt{\text{MSE}}} \right)
\]

\[
= 20 \cdot \log_{10} (\text{MAX}_I) - 10 \cdot \log_{10} (\text{MSE})
\]
Fig 3 Comparison of PSNR values for both initial and final stage

The above figure 3 clearly depicts the comparison of the proposed algorithm at both initial and final stage on the basis of PSNR values. The above results quiet familiarize that the values at final stage is quiet reliable and optimist as compare to value as at initial stage. Higher will be the PSNR values higher will be the result accurate.

Fig 4 Comparison of RMSE values for both initial and final stage

The above Fig 4 clearly depicts the comparison of the proposed algorithm at both initial and final stage on the basis of RMSE values. The above results quiet familiarize that the values at final stage is quiet reliable and optimist as compare to value as at initial stage. Smaller will be the RMSE values higher will be the result accurate.

V. CONCLUSION & FUTURE SCOPE

In this paper, the proposed work stressed towards the phase based binarization and segmentation of Ancient Document Images. This paper has focused on the degraded document binarization technique. Document binarization is a key application of vision processing. The main objective of this paper is to evaluating the short comings of algorithms for degraded image binarization. It has been found
that each technique has its own benefits and limitations; no technique is best for every case. The proposed performance and comparison was made over each images stored in the set of data set of Ancient Document images. Here, the dissertation taken two metrics namely PSNR and RMSE. The results clearly depicts that the value of PSNR for all the images stored in the set of dataset is quiet high at final stage in compare to the value at initial stage which is quiet good, reliable and scalable for data storage and removes the data redundancy. The proposed result is quiet scalable and optimistic one for set of datasets and founds that there is higher bit of result at final stage as compare to initial stage.

Furthermore, there is need to improve the results more by introducing more feature extraction phase and also work over hand written document images. There is also need to use of some more technique and maybe comparison of different technique. As a future scope, one may also choose some more parameters to compare the results

REFERENCES

Ms. Ritika Sood is pursuing her M.Tech. in ECE from CGC-COE, Landran, Punjab, INDIA. Her field of interest is in Image Processing based system development and integration.