2 CLEANING WITH MODIFIED ITERATIVE GLOBAL THRESHOLD (MIGT) AND SEGMENTATION ON NOISY/NON-NOISY DOCUMENTS
2.1 Introduction

In an automated processing system for a historical document, binarization is meant for cleaning the noisy documents thus enabling the smooth functioning of later stages. The line segmentation method’s performance [68] is known to be significantly influenced regarding its segmentation and recognition accuracy. Followed by the work reported in [98] for optical character recognition of Telugu script, an experimental analysis is made for a method depending on text segmentation to isolate patterns having high noise levels. Later binarization is applied to the noisy documents to show the difference of performing segmentation on noisy and noise-less document images.

2.2 General Flow Chart

Although, OCR is known to involve two stages of it being processed, a noisy Telugu text document image generally considers horizontal projection profile for its segmentation into lines. However, character segmentation is developed, through selecting proper threshold, with the help of a vertical profile.

The adopted technique for the segmentation of an ancient noisy Telugu document is presented as flow chart in Figure 2.1 with the details of the stages of cleaning, etc.

Nevertheless an efficiency parameter is designed as a benchmark to estimate the performance of the present model. It is a serial processing system. After completion of the phase-1, the system processes phase-2.
CLEANING WITH MODIFIED ITERATIVE GLOBAL THRESHOLD (MIGT) AND SEGMENTATION ON NOISY/NON-NOISY DOCUMENTS

Figure 2-1 Flow Chart for Proposed Methodology

The first phase involving the segmentation of noisy documents. Binarization is an importance in document image analysis. Telugu script segmentation into resourceful units is quite complex due to the cursive nature and hence became a challenging objective.
A pure gray-scale image provides foreground/background information of the image. Black and white is designated as '0' and '255' respectively. Segmentation specifically depends on the characteristic noise. In general, there would be a change in background and foreground pixel intensity and is treated as noise. Along with the foreground information, it contains gaps between lines and characters. It causes pixel intensity changes because of presence of noise.

Effective segmentation can be achieved using gray-scale dependent profile information analysis. Peak width of intensity provides fundamental information towards lines and characters separation. The peak width gradually decreases between lines and characters. First order Gaussian kernel with sigma-3 is used for convolution of horizontal profile information and is shown in Figure 2.2.

![Text Sample Horizontal Profile](image)

**Figure 2-2 Text Sample Horizontal Profile**

Character segmentation is done by adopting a characteristic threshold generated from the vertical profile intensities. A representative vertical profile is presented in Figure 2.3. Significant
changes in the characteristics of vertical profile are found in contrast to that for the horizontal profile information. Finite width peaks are identified in the horizontal profile and are not clearly identified in a vertical profile. Further, the peak appears with non-uniform intensities.

![Figure 2-3 Text Line Vertical Profile](image)

The 2nd phase processing deals with segmentation of cleaned (i.e., for the case of noise free documents) Telugu text documents. A Modified Iterative Global Threshold technique is used during the cleaning of documents followed by segmenting it into lines and words by using horizontal profile, while character segmentation is carried out by using the vertical profile. In cases of any touching characters, they are identified by using Drop fall algorithm, which in turns improves the segmentation accuracy.

### 2.3 Algorithm for Segmentation of Noisy Documents

An algorithm comprised of a series of sequential steps implemented undergoes various stages such as noisy document extraction; profile (horizontal) identification; convolution of Gaussian kernel with horizontal profile; identification of peaks;
CLEANING WITH MODIFIED ITERATIVE GLOBAL THRESHOLD (MIGT) AND SEGMENTATION ON NOISY/NON-NOISY DOCUMENTS

profile (vertical) identification of line; defining threshold intensity and peaks identification for segmentation of characters.

A document image in noisy environment is given by a vector $I(n,m)$ in which 'n' and 'm' are number of lines and columns. Horizontal profile $I(n,m)$ is identified by sum of all pixel intensities perpendicular to Y-axis (HP with a size 'n') is:

$$HP[i] = \sum_{j=1}^{m} I(i,j)$$

where,

$HP[i]$ is the Horizontal Profile

$I(i,j)$ is the image

'i' number of rows of an image $I(i,j)$

'j' number of columns of an image $I(I,j)$.

For a line separation direct horizontal profile, peaks and valleys identification are quite complex due to the large value of each pixel and hence finding a threshold is a complex task. The profile is convolved with Gaussian first order kernel and is given by Eq.-2.1.

![Gaussian Kernel with Sigma-3](image)

For perfect segmentation of lines, first order differentiation of Gaussian kernel is suitable. Gaussian kernel information deciphered

33
CLEANING WITH MODIFIED ITERATIVE GLOBAL THRESHOLD (MIGT) AND SEGMENTATION ON NOISY/NON-NOISY DOCUMENTS

is one and similar for higher orders but it gives a sort of complex computation. Gaussian kernel width for sigma-3 is illustrated in Figure 2.4 and suitable for line segmentation.

Segmentation efficiency during the processing is found to posses varied order value and sigma value. Convolution is a linear operation under Gaussian kernel represented by:

\[ G = \left[ \frac{1}{2\pi \sigma^2} \right] e^{\left[ \frac{-x^2}{2\sigma^2} \right]} \]

where,

\( G \) is the Gaussian probability density \\
\( X \) is the mean value \\
\( \sigma \) width of the Gaussian kernel (Standard Deviation)

During convolution process, Gaussian kernel degree of shift linearly varies with horizontal profile. Hence convolution would be used to show profile randomness and gives representative zero-crossing smoothing curve. It is represented as in Figure 2.5 when convolved with profile.

![Figure 2.5 Convolution Outcome between Gaussian Kernel and Horizontal Profile](image-url)
The peaks above zero are treated as the gaps between lines. Then the line segmentation is performed based on above information. Gaussian kernel and respective convolving profile with are illustrated in Figures 2.4 and 2.5.

The resultant equation after convolving the profile with Gaussian kernel is represented by Eq.-2.3

\[ C = \int G * HP \, dt \]  \hspace{1cm} (2.3)

where,

\( G \) is the Gaussian probability density

\( HP \) is the Horizontal Profile

From Figure 2.5, it is noticed that Gaussian kernel represents a smooth profile with a space among immediate peaks. Positive peak and negative peak represent space among lines and data. Thus the extraction of lines from text is a simpler operation.

Line vertical profile is generated using computation of sum of pixel intensities perpendicular to X-axis (VP) of line width ‘d’ which is defined by Eq.2.4.

\[ VP[j] = \sum_{i=1}^{n} I(i,j) \]  \hspace{1cm} (2.4)

where,

\( VP[j] \) is the Vertical Profile

\( I(i,j) \) is the image

‘\( i \)’ number of rows of an image \( I(i,j) \)

‘\( j \)’ number of columns of an image \( I(I,j) \)

Adoption of a similar Gaussian kernel for segmentation of character gives a smoother profile having peaks with non-
uniformity. Non-uniform peaks recognition with a specific intensity value contains certain levels of difficulty and hence fixing a threshold for character segmentation plays a crucial role. Fixing a suitable threshold from vertical profile value would be a prerequisite condition. Experimentation of threshold value is through their maxima and minima values of vertical profile as given below.

\[ I_{Th} = (VP)_{max} - \frac{(VP)_{max} - (VP)_{min}}{3.6} \]  

where,

- \( I_{Th} \) is the Threshold
- \((VP)_{max}\) is the Maximum value of Vertical Profile
- \((VP)_{min}\) is the Minimum value of Vertical Profile

In the second phase of segmentation, the ancient documents are cleaned by using Modified Iterative Global Threshold Algorithm and the text document is segmented into lines, words and characters.

### 2.4 Cleaning of Noisy Documents

Binarization is one of the important steps of processing implemented for documents which are exposed to image analysis. It consists of labeling of each pixel in the image classified as foreground and background. It provides a proper distinction between background and foreground. In this thesis the above said algorithm is used for cleaning the documents.

The interesting aspects of a historical document image analysis created several challenges. Degraded conditions like; bleed through, ink stains, torn pages, etc., motivated research for their binarization. Further, enhancement of efficiency of relevant
CLEANING WITH MODIFIED ITERATIVE GLOBAL THRESHOLD (MIGT) AND SEGMENTATION ON NOISY/NON-NOISY DOCUMENTS

algorithms is a challenging task. Binarization is referred as first stage in various current systems of image processing. OCR of cultural heritage documents includes the collection of digitized images. Such image's antique importance is known on the internet for manual annotations. Having original documents at home gives an opportunity to handle old material with historical importance, into a price of attention to pay. Frequent handling of these documents results in the physical wear and tear thus leading to eventual damage of the document. In addition, the documents so treated are vulnerable to the damage caused by fluctuating environments. In this direction of keeping a distinction of ancient documents, their processing with OCR plays an important role for removing the background noise and improving its readability.

Figure 2-6 Flowchart of the Proposed Model

Two simple and practical Binarization [16] techniques are known to be adopted viz. fixing the Global and Local threshold values. The global threshold [30] defines a global value for all the
pixel intensities of the image to separate them as text background. This method fails to remove non-uniform noise. Local threshold gives an adaptive solution for the images with different background intensities. However the threshold value of the later method varies according to the properties of the local region. There are many general purpose Binarization methods capable of dealing with any type of document image, especially with complex background. All these methods fall under local or adaptive [23,33] thresholding techniques.

A generalized method is proposed for degraded documents cleaning using modified Iterative Global Threshold algorithm. A simple approach followed for separating information of objects from foreground. It involves a global threshold computation and two distinct clusters will be separated. An iterative methodology is adopted to handle various degradation conditions of documents. Intermediate tones are shifted at each iteration towards the background thus improving the efficiency of bringing a separation of foreground. It is a known suitability for the documents of non-uniform noise distribution. Presently a technique (Figure 2.6) is evolved for binarization of ancient documents which relies on the clustering of pixels. In this technique, gray level data undergoes a clustering analysis with the number of clusters being set always to two. These two clusters correspond to two peaks of a histogram. Here mid-point of pixels is computed. The flowchart of proposed model developed currently with details of binarization and benchmark for its efficiency is illustrated in Figure 2.6.

The conventional approach in a Global thresholding [41] technique targets to find a unique threshold value to eliminate all the pixels that represent the background of image, and preserves other forms of foreground of image. Usually many real world images
possess complex backgrounds and/or weak image foregrounds. In realistic situations, some foreground pixels accompany with the gray values very close to that of background pixels. Hence, it would be difficult to determine a single threshold value that completely separates the object information from the background. Similar argument could be extended even for local thresholding technique i.e., connecting pixel by pixel, or region by region. However, for the presently proposed algorithm, after each operation of thresholding image equalization process is carried out. Thus, while evaluating the relative importance of respective pixel intensity in the background image is evaluated. Thus, evaluation of the new threshold performs the process of elimination of background. This process will be repeated (iterated) till a sensitive threshold value is achieved.

### 2.5 Modified IGT Algorithm

Series of sequential steps are necessarily involved in executing the Modified IGT algorithm, especially for noisy documents. They include extraction of degraded (noisy) document, conversion of noisy document into gray-scale image, averaging its intensity of background plus that of the object, intensity shifting of pixels in the image (towards background), equalization of the image (related to that of foreground object intensity) on background information, determination of the average intensity of the resultant image and evaluation of the threshold (between the iterative average intensities).

A gray-scale image of the noisy document is generally represented by

$$ I(x, y) = S, S \in s[0, 1] \quad \text{----------(2.6)} $$

Here $x$ is the horizontal and $y$ is the vertical coordinates of the image $I(x,y)$. 
S takes any value between 0 and 1, while \( S=1 \) stands for white and \( S=0 \) stands for black.

Intermediate tones are shifted to the background in the proposed algorithm. In general, any document image is known to include only a few pixels of useful (foreground) information against the image size which is a combination of foreground and background. Rarely the amount of object information exceeds 10% of the total pixels in the document. In the wake of this as fact of advantage, the average value of the pixels will be determined mainly by the background information. There are two parts in the proposed Modified IGT algorithm. In the first part, the level shifting of the pixels for an image is performed, while, the relative importance of pixels is determined with respect to object information in the second part of the algorithm. After each iteration, some of pixels would be moved from the fuzzy region to the background. The iteration process will continue till the following criterion is satisfied by the Eq.2.7

\[
|T_i - T_{i-1}| = t
\]

where,
- \( T_i \) is the threshold used in \( i^{th} \) iteration and
- \( T_{i-1} \) is the threshold before the \( i^{th} \) iteration
- \( t \) is the sensitivity parameter of threshold

The threshold \( T_i \) is the average intensity of background + object for an MxN document image, expressed by

\[
T_i = \frac{\sum_{x=1}^{M} \sum_{y=1}^{N} I_i(x,y)}{M \times N}
\]

where,
CLEANING WITH MODIFIED ITERATIVE GLOBAL THRESHOLD (MIGT) AND SEGMENTATION ON NOISY/NON-NOISY DOCUMENTS

\[ T_i \] is the average intensity of background + object
\[ I_i(x,y) \] is the gray-scale image
\[ M \] is the number of rows and
\[ N \] is the number of columns of an image.

Depending on the type of object represented in gray level, the following assumptions are made; '0' is attributed for black color as the text, 255 is the highest attributed luminance of document (or conversely the foreground as white and the background as black). A threshold is computed as initial separating point of background and foreground layers. In each iteration the threshold value of the image would catch the background cluster. When the background is white, the threshold value is shifted towards background and the converse operation (Figures 2.12 and 2.13) is performed.

![Figure 2-7 Threshold Shift in each Iteration for Black Foreground](image-url)
During the process of cleaning a cluster of pixels, which are situated in the fuzzy region (i.e. belongs to either foreground or background) they are shifted using the equation given by the Eq.2.9 towards background cluster when the image background is white.

On the other hand the image background is attributed black while the fuzzy pixels are shifted towards foreground as given by Eq.2.10. As all iterations are performed for the level shifting of the pixels, approximately 85% of pixels are identified so as to belong to fuzzy region, and they would be shifted to background or foreground.

\[ I_s(x, y) = [L - T_i] + I_i(x, y) \quad \text{----------}(2.9) \]

(or)

\[ I_s(x, y) = I_i(x, y) - T_i \quad \text{----------}(2.10) \]

where,

\( T_i \) is the average intensity of background + object

\( L \) is a variable to define the maximum or minimum luminance value.

\( I_s(x,y) \) is the resultant image after 1\textsuperscript{st} iteration
CLEANING WITH MODIFIED ITERATIVE GLOBAL THRESHOLD (MIGT) AND SEGMENTATION ON NOISY/NON-NOISY DOCUMENTS

$I_i(x,y)$ is the gray scale image

After the process of level shifting of pixels is completed, the left over pixels in the fuzzy region would undergoing for equalization process to determine the influence of foreground object intensity on background information for white background images as given by Eq.2.11. For black background images as given by Eq.2.12, where “k” is the sensitivity parameter. The amount of intermediate tones shifted to background or foreground cluster depends on the value of “k”.

$$I_r(x,y) = I_s(x,y) - k \left[ \frac{L-I_s(x,y)}{L-E_i} \right] \quad \text{(2.11)}$$

(or)

$$I_r(x,y) = I_s(x,y) + k \left[ \frac{L-I_s(x,y)}{L-E_i} \right] \quad \text{(2.12)}$$

where,

$K$ is the sensitivity parameter

$L$ is a variable to represent the maximum or minimum luminance value.

$E_i$ is the minimum/maximum pixel intensity value.

$I_s(x,y)$ is the equalized image

$I_r(x,y)$ is the resultant image after image equalization during $i^{th}$ iteration

$I_r(x,y)$ is given by the Eq.2.11&2.12 and $E_i$ is the minimum (for black background images) or maximum (for white background images) pixel intensity value in the image $I_s(x,y)$ during $i$-th iteration just before image equalization.

By substituting the Eq.2.9 in Eq.2.11 and Eq.2.10 in Eq.2.12. The resultants for $I_r(x,y)$ are given
CLEANING WITH MODIFIED ITERATIVE GLOBAL THRESHOLD (MIGT) AND SEGMENTATION ON NOISY/NON-NOISY DOCUMENTS

\[
I_r(x, y) = L - T_i + I_i(x, y) - k \left[ \frac{T_i - I_i(x, y)}{L - E_i} \right] \quad \text{(2.13)}
\]

(or)

\[
I_r(x, y) = I_i(x, y) - T_i + k \left[ \frac{T_i - I_i(x, y)}{L - E_i} \right] \quad \text{(2.14)}
\]

where,

\( K \) is the sensitivity parameter

\( L \) is a variable to represent the maximum or minimum luminance value

\( T_i \) is the average intensity of background + object

\( E_i \) is the minimum or maximum pixel intensity value

\( I_i(x, y) \) is the gray scale image

\( I_r(x, y) \) is the resultant image after image equalization during \( i^{th} \) iteration

### 2.6 Conclusion

Modified Iterative Global Thresholding is proposed. Binarization is tested with clustered approach through most likely background data estimation. Average intensity in every iteration is used as clusters’ mid-point. For histogram compand, equalization is done for remaining pixels in the next stage. Successive threshold sensitivity decides the number of iterations. The proposed method is proved as an effective approach targeting camera-captured stone carvings as well as historical document images. However, it is observed that further improvement is necessary for applying on palm leaf manuscripts. Also, segmentation is performed on both noisy and cleaned documents and compared against various factors.