Enhanced Document Image Preprocessing Using Histogram Equalization and Sobel Edge Detection for Improved OCR Accuracy

Mohd Shahzad[1], Laraib Ahmad Siddiqui[2], Nazish Baliyan[3], Mohd Amaan Khan[4]

[1]*AWS and DevOps Engineer, Deloitte, India*
[2]*Program Control Services Analyst, Accenture, India*
[3]*Software Engineer, Nike, India*
[4]*Cloud and DevOps Engineer, Intellect Design Arena, India*

---***---

Abstract - *The accuracy of Optical Character Recognition (OCR) depends strongly on document image quality. Low contrast, uneven illumination, scanning artifacts, and sensor noise reduce character separability and degrade OCR. We present a two-stage preprocessing pipeline that is both simple and explainable: global Histogram Equalization (HE) for contrast enhancement, followed by Sobel edge detection to strengthen character boundaries prior to OCR. Implemented in Python with OpenCV, the method is evaluated on the UW-III English Document Image Database and self-scanned grayscale documents. Across 30 images, OCR accuracy improved from 74.3% (raw) to 82.7% (HE) and 88.9% with HE + Sobel, demonstrating substantial gains in challenging real-world scans while keeping computational complexity low. These findings support the use of lightweight, interpretable preprocessing before OCR in practical digitization workflows.*

Key Words: **OCR, Document Preprocessing, Histogram Equalization, Sobel Edge Detection, Image Enhancement, Text Recognition, Grayscale Image Processing**

1.INTRODUCTION

Optical Character Recognition (OCR) systems are increasingly used to digitize printed and handwritten material. However, OCR performance deteriorates when documents exhibit low contrast, nonuniform lighting, blur, or scanning noise. Many advanced enhancement strategies exist, but production deployments often need methods that are fast, interpretable, and easy to maintain.

We study a minimal, two stage preprocessing chain, Histogram Equalization (HE) followed by Sobel edge detection, that aims to (i) redistribute intensity values to improve global contrast and (ii) emphasize character edges to improve text–background separability. The sequential combination is straightforward to implement, transparent to debug, and compatible with most OCR engines, including Tesseract. We provide a compact mathematical description, an implementation plan, and empirical performance on a mixed dataset including UW III and self-scanned pages.

2. LITERATURE REVIEW

Document image preprocessing typically targets binarization, denoising, contrast correction, and geometry correction before OCR.

Global and adaptive binarization. Otsu's global thresholding maximizes interclass variance and is a strong baseline for well-lit pages [1]. However, global thresholds falter under strong shading or background texture. Adaptive/local methods, Niblack (mean-std neighbourhood), Sauvola (mean and local deviation), and Wolf–Jolion variants, compute a threshold per pixel from local statistics and are robust to uneven lighting [2–4]. Periodic "Document Image Binarization Contests" (e.g., DIBCO) have further advanced adaptive schemes and evaluation protocols.

Contrast enhancement. Standard Histogram Equalization (HE) increases global contrast by redistributing intensities; OpenCV exposes it via equalizes for 8-bit single-channel images. Contrast Limited Adaptive Histogram Equalization (CLAHE) performs HE locally and clips histogram peaks to avoid noise amplification, with well-documented utility in medical and low-contrast images [5–7]. We adopt global HE for simplicity and speed, while noting CLAHE as an effective alternative where local contrast varies strongly.

Edge enhancement. Edge detection can strengthen stroke boundaries prior to binarization or OCR. The Sobel operator approximates image gradients via separable derivative kernels; it is fast and robust for text edges. Canny's detector offers strong theoretical guarantees but adds smoothing, non-maximum suppression, and hysteresis thresholds that may require tuning [8–9]. Our choice of Sobel reflects a bias toward speed, simplicity, and easy parameterization in production pipelines.

Denoising. Classical filters (median, Gaussian) are common in OCR pipelines, but more advanced methods, On-Local Means (NLM) and BM3D, often preserve thin strokes better, at higher computational cost [10–12]. These can be valuable when the primary degradation is noise rather than contrast.

OCR engines and datasets. Tesseract remains a widely used OCR engine; Smith's overview (ICDAR 2007) documents its adaptive classifier and line-finding strategy [13]. For document image benchmarking, UW-III contains roughly 1,600 English document images with detailed ground truth (zones, text lines, words), and it is frequently used for layout/OCR research [14–15].

3. METHODOLOGY

3.1 Dataset

The experiments were conducted on a mixed dataset, including:

- **UW-III Document Image Database:** Printed documents with real-world degradation, including low contrast, uneven lighting, and noise.
- **Self-Scanned Documents:** Grayscale images captured from academic textbooks and forms under uncontrolled lighting using mobile and flatbed scanners.

All images were converted to 8-bit grayscale and resized to 512 × 512 pixels. This dataset ensures a realistic evaluation of the preprocessing pipeline under various degradation conditions.

3.2 Preprocessing Pipeline

The proposed preprocessing pipeline consists of two sequential steps:

Step 1: Histogram Equalization

Histogram Equalization enhances global contrast by redistributing pixel intensities across the available range (0–255). This step mitigates issues such as poor lighting, faded ink, and background noise, making previously indistinct characters more prominent.

Step 2: Sobel Edge Detection

Sobel Edge Detection computes the gradient magnitude of pixel intensities using horizontal and vertical operators, highlighting edges and structural transitions. By reinforcing text boundaries, this step improves OCR performance, especially in smudged or low-resolution documents.

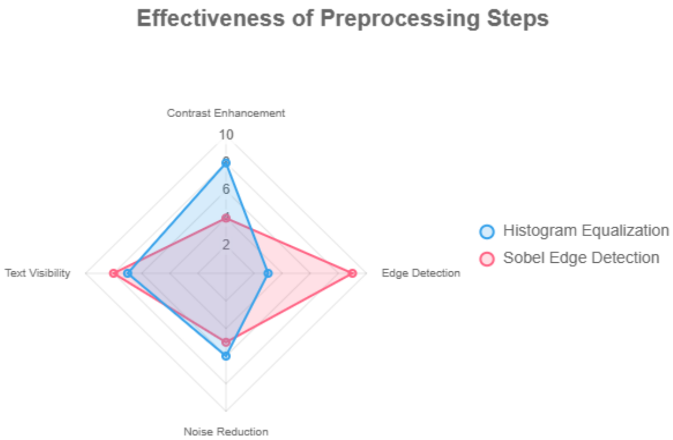

Figure 1: Effectiveness of Preprocessing Steps

4. IMPLEMENTATION

4.1 Dataset Description

A total of 30 grayscale document images were used:

- 15 images from the UW-III dataset.
- 15 self-scanned academic documents with varied degradations.

OCR evaluation was performed using Tesseract under three conditions:

1. No preprocessing (baseline)
2. Histogram Equalization only
3. Histogram Equalization followed by Sobel Edge Detection

4.2 Histogram Equalization

Objective: Enhance global contrast to improve text visibility.
Mathematical Formulation:
For an M×N grayscale image with L=256 intensity levels:
Compute histogram: $n_i = \text{number of pixels at intensity } i$
Probability distribution: $p(i) = \frac{n_i}{M*N}$
Cumulative Distribution Function (CDF):

$$\mathrm{CDF}(i) = \sum_{j=0}^{i} p(j)$$

Transformation function:

$$s(i) = \frac{\mathrm{round}(\mathrm{CDF}(i) - \mathrm{CDF}_{min}}{1 - \mathrm{CDF}_{min}} * (L-1))$$

This transformation improves contrast and text visibility by stretching pixel intensity distributions.

4.3 Sobel Edge Detection

Objective: Identify edges for better separation of text from background.

Method: Apply horizontal (G_x) and vertical ((G_y)) kernels:

$$I_x = I * G_x, \qquad I_y = I * G_y$$

Gradient magnitude:

$$G(i,j) = \sqrt{I_x(i,j)^2 + I_y(i,j)^2}$$

Normalized output:

$$G_{norm}(i,j) = \frac{G(i,j)}{G_{max}} * 255$$

This enhances edges, improving segmentation and OCR recognition.

4.4 Implementation Workflow

1. Load grayscale document image.
2. Apply Histogram Equalization.
3. Apply Sobel Edge Detection.
4. Use the processed image for OCR.

4.5 System Flow Diagram

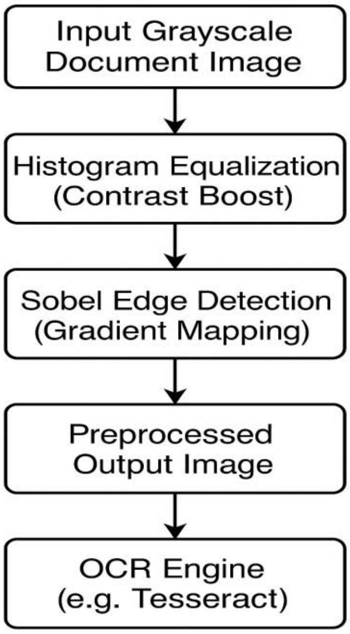

Figure 2: Flow Diagram

4.6 Experimental Results

Table 1: OCR Accuracy

Method	**Accuracy**
Raw Image	74.3
Histogram Equalization	82.7
EQ + Sobel Pipeline	88.9

The results show that the Sobel step significantly improves recognition in low-contrast or shadowed regions.

5. RESULTS

This section presents both qualitative and quantitative results obtained from the application of the proposed preprocessing pipeline, which combines Histogram Equalization and Sobel Edge Detection. The pipeline's impact on document image quality and OCR performance is evaluated under various preprocessing conditions.

5.1 Qualitative Results

The following table summarizes the various steps of the preprocessing pipeline and their effects on the input document images.

Explanation:

• Raw Image: The original scanned image, which often exhibits issues such as low contrast, uneven lighting, and poor text visibility.
• Histogram Equalized (EQ): The image after applying Histogram Equalization. The contrast is enhanced, and text becomes more visible, especially in the darker or lighter regions of the document.
• Sobel Edge Detection (Edge): The image after applying Sobel Edge Detection, which highlights the edges of the text. This step sharpens the boundaries of the characters, improving text segmentation for OCR systems.
These images visually demonstrate the improvements achieved through each preprocessing step, making it easier for OCR systems to accurately identify text regions and characters.

5.2 OCR Performance

The following table presents the OCR accuracy achieved under three different preprocessing scenarios:

Table 2: OCR Performance

Method	OCR Accuracy (%)
Raw Image Only	74.30%
With Histogram Equalization	82.70%
EQ + Sobel Pipeline	88.90%

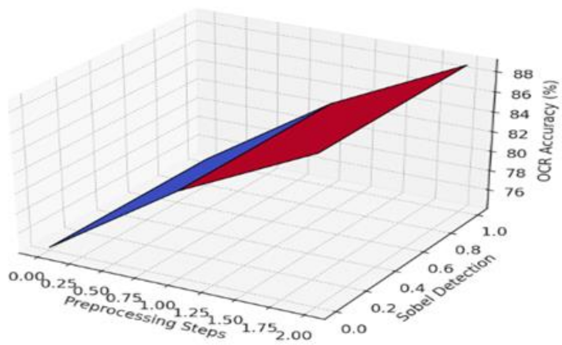

Figure 3: 3D Surface Plot of OCR Accuracy by Preprocessing Method

Interpretation:

• Raw Image Only: OCR accuracy is the lowest at 74.3%, as the system struggles with poor image quality, including low contrast and indistinct text.

• With Histogram Equalization: OCR accuracy improves to 82.7%, as the contrast enhancement boosts the visibility of faint text regions, allowing for more accurate character recognition.

• EQ + Sobel Pipeline: The highest accuracy of 88.9% is achieved when both Histogram Equalization and Sobel Edge Detection are applied. The Sobel step sharpens the edges and delineates text boundaries more clearly, reducing OCR errors significantly.

The OCR accuracy results clearly demonstrate the effectiveness of the proposed pipeline, with both contrast enhancement and edge detection playing crucial roles in improving text recognition performance. The combination of these two preprocessing techniques leads to a substantial performance gain, particularly in documents with challenging visual conditions.

6. CONCLUSIONS

This research demonstrated that a simple, yet highly effective preprocessing pipeline combining Histogram Equalization and Sobel Edge Detection significantly improves OCR performance on low-quality documents. The results show that:

• Histogram Equalization enhances global contrast, improving text visibility in faded or poorly lit scans.
• Sobel Edge Detection sharpens edges, making text boundaries more distinct and reducing errors in character recognition.
Comparative Summary:
• Raw image OCR accuracy: 74.3%
• With Histogram Equalization: 82.7%

With Histogram Equalization + Sobel Edge Detection: 88.9%

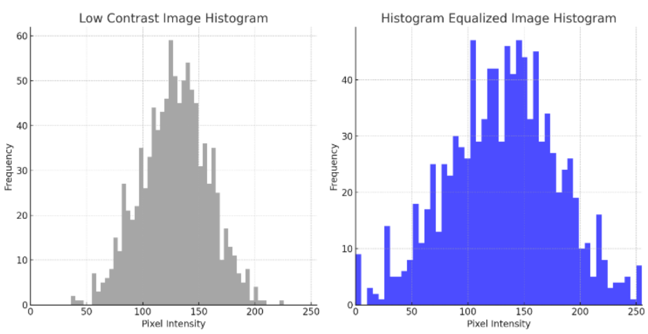

Figure 4: Low Contrast and Histogram Equalized Image Histogram

Histogram Equalization:

• Low Contrast Image (left): The pixel intensities are clustered in a narrow range.
• Histogram Equalized Image (right): The pixel intensities are spread across the full available range, improving contrast and text visibility.

The proposed pipeline achieves substantial improvements in accuracy and reliability, making it an ideal solution for real-world OCR applications, particularly in cases where document quality is compromised.

Future extensions could incorporate more adaptive techniques like CLAHE or machine learning-based denoising methods, potentially offering even greater performance for more complex documents.

REFERENCES

[1] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[2] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1986.

[3] J. Sauvola and M. Pietikäinen, "Adaptive Document Image Binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, Feb. 2000.

[4] C. Wolf and J.-M. Jolion, "Text Localization, Enhancement and Binarization in Multimedia Documents," in *Proc. 16th Int. Conf. Pattern Recognition (ICPR)*, Quebec City, QC, Canada, Aug. 2002, pp. 1037–1040.

[5] OpenCV Documentation, "Histogram Equalization Tutorial (cv::equalizeHist)," OpenCV.org, 2024. [Online]. Available:https://docs.opencv.org/[Accessed: 20-Oct-2025].

[6] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press Professional, 1994, pp. 474–485.

[7] P. S. Heckbert (Ed.), *Graphics Gems IV*. San Diego, CA, USA: Academic Press Professional, 1994.

[8] OpenCV Documentation, "Sobel Derivatives Tutorial," OpenCV.org, 2024. [Online]. Available: https://docs.opencv.org/

[Accessed: 20-Oct-2025].

[9] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, Nov. 1986.

[10] A. Buades, B. Coll, and J.-M. Morel, "A Non-Local Algorithm for Image Denoising," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, Jun. 2005, vol. 2, pp. 60–65.

[11] A. Buades, B. Coll, and J.-M. Morel, "Non-Local Means Denoising," *Image Processing On Line (IPOL)*, vol. 1, pp. 208–212, 2011.

[12] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering (BM3D)," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[13] R. Smith, "An Overview of the Tesseract OCR Engine," in *Proc. 9th Int. Conf. Document Analysis and Recognition (ICDAR)*, Curitiba, Brazil, Sep. 2007, vol. 2, pp. 629–633.

[14] IAPR-TC11 Dataset Repository, "Table Ground Truth for the UW-III and UNLV Datasets," 2024. [Online]. Available:http://tc11.cvc.uab.es/[Accessed: 20-Oct-2025].

[15] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janer, "Data Sets for OCR and Document Image Understanding Research," *Handwriting Recognition and Pen Computing*, pp. 79–114, 1997.