Newspaper OCR Pre-Processing & OCRmyPDF Workflow

Overview:

This project processes scanned newspaper PDFs that contain images. The code first enhances the image quality of each page while preserving the original color. It then applies OCR to add an invisible text layer to the PDF. This makes the final document searchable and selectable without changing its visual appearance.

Features:

- High-resolution conversion: Each PDF page is converted to an image at 300 DPI to ensure high quality.
- Color-preserving pre-processing: Instead of converting images to grayscale, the code adjusts contrast and brightness and applies a color-preserving denoising filter.
- Reassembly: The processed color images are reassembled into a new PDF.
- OCR integration: OCRmyPDF adds an invisible text layer, with options to deskew and optimize the file size.
- Searchability: The resulting PDF contains a text layer that allows users to select and search text within the newspaper.

Workflow:

1. PDF to Image Conversion:
   - The code uses a tool (pdf2image) to convert each page of the original newspaper PDF into a high-resolution image.
2. Color-Preserving Image Pre-Processing:
   - Each image is processed with OpenCV.
   - Contrast and brightness are enhanced using cv2.convertScaleAbs.
   - A color-preserving denoising filter (cv2.fastNlMeansDenoisingColored) is applied to reduce noise while keeping the colors intact.
3. Reassembly of the PDF:
   - The enhanced images are reassembled into a new PDF using the Pillow library.
4. OCR Application:
   - OCRmyPDF is then used to process the preprocessed PDF.
   - It applies deskewing and optimizes file size.
   - It adds an invisible text layer using Tesseract (with a specific page segmentation mode), making the PDF searchable.

Requirements:

Python 3.x

Required Python packages:

- ocrmypdf
- pdf2image
- opencv-python
- numpy
- Pillow

Poppler utilities (for pdf2image):

- On macOS: Install via Homebrew (brew install poppler)
- On Linux/Ubuntu: Install via sudo apt-get install poppler-utils

How to Use:

1. Place your original newspaper PDF in the designated location.
2. Run the script. It will:
   - Convert each page to a high-resolution image.
   - Process each image to enhance quality while preserving color.
   - Reassemble the processed images into a new PDF.
   - Run OCR on the new PDF to add a searchable text layer.
3. The output will include:
   - A preprocessed PDF with enhanced, color-preserved images.
   - A final PDF with an invisible text layer that makes the document searchable and selectable.

Summary:

This workflow is designed to improve OCR accuracy for newspaper PDFs by using color-preserving image enhancements before OCR. The final PDF retains its original visual quality while becoming fully searchable, making text extraction and review much easier.

Output

Before

Ernest A. Rice, 51, Local Accountant, Dies

Ernest A. Rice, 51, 567 Tingler Road, died Tuesday at Methodist Hospital in Indianapolis following an illness of three months.

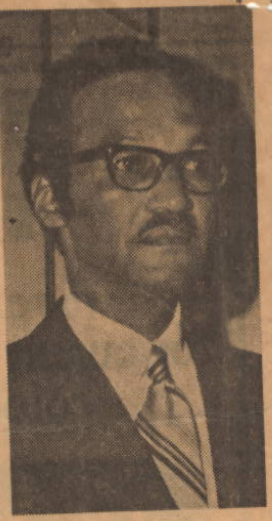
He taught night school accounting classes at Richmond High School seven years and was employed as an accountant at Philco-Ford in Connersville. He also maintained a public accountant and tax consultant office at 421 South Fifth St.

He had been employed at Richmond Supply Corp. for 21 years and had served as executive vice president of the firm.

A native of Richmond, he was graduated from Morton High School in 1939, and was a graduate of West Virginia State College and took postgraduate work at Indiana University extension. He also was a veteran of World War II.

A member of Bethel A.M.E. Church and the National Accounting Association, he was on the board of directors of the Boys' Club, Onyx Lodge, American Legion, Esquire Club and Alpha Psi Alpha Fraternity.

Survivors include his widow, Bette; one son, Alan, at home; one stepson, Michael Spicer of Minneapolis, Minn.; his mother and stepfather, Mr. and Mrs. John Dillard of Richmond; one sister, Mrs. Elizabeth Owens of Cleveland, Ohio; four brothers, Emmett and Harry of Richmond, Howard of Cincinnati, Ohio, and Robert of Dayton, Ohio; two aunts and nieces and nephews.



Ernest A. Rice

Services for Mr. Rice will be at 1 p.m. Saturday at the Bethel A.M.E. Church with Rev. J. P. Henning officiating. Burial will be in Earlham Cemetery. Friends may call from 2 to 9 p.m. Friday at the Patterson Funeral Home where memorial services will begin at 7:30 p.m. Friday, opened by the Onyx Lodge No. 479, and followed by the Moore-Irvin Post No. 359 of the American Legion.

After

Ernest A. Rice, 51, Local Accountant, Dies

Ernest A. Rice, 51, 567 Tingler Road, died Tuesday at Methodist Hospital in Indianapolis following an illness of three months.

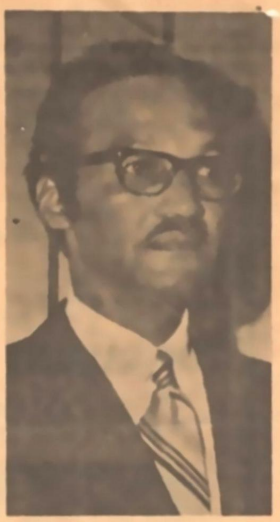
He taught night school accounting classes at Richmond High School seven years and was employed as an accountant at Philco-Ford in Connersville. He also maintained a public accountant and tax consultant office at 421 South Fifth St.

He had been employed at Richmond Supply Corp. for 21 years and had served as executive vice president of the firm.

A native of Richmond, he was graduated from Morton High School in 1939, and was a graduate of West Virginia State College and took postgraduate work at Indiana University extension. He also was a veteran of World War II.

A member of Bethel A.M.E. Church and the National Accounting Association, he was on the board of directors of the Boys' Club, Onyx Lodge, American Legion, Esquire Club and Alpha Psi Alpha Fraternity.

Survivors include his widow, Bette; one son, Alan, at home; one stepson, Michael Spicer of Minneapolis, Minn.; his mother and stepfather, Mr. and Mrs. John Dillard of Richmond; one sister, Mrs. Elizabeth Owens of Cleveland, Ohio; four brothers, Emmett and Harry of Richmond, Howard of Cincinnati, Ohio, and Robert of Dayton, Ohio; two aunts and nieces and nephews.



Ernest A. Rice

Services for Mr. Rice will be at 1 p.m. Saturday at the Bethel A.M.E. Church with Rev. J. P. Henning officiating. Burial will be in Earlham Cemetery. Friends may call from 2 to 9 p.m. Friday at the Patterson Funeral Home where memorial services will begin at 7:30 p.m. Friday, opened by the Onyx Lodge No. 479, and followed by the Moore-Irvin Post No. 359 of the American Legion.