SUMMER INTERNSHIP REPORT

TOPIC: DENOISING SCANNED INVOICE IMAGES

COMPANY: EMAGIA CORPORATION

SUBMITTED BY: SAI CHANDANA - BE(AIML – 3rd yr)

COLLEGE OF ENGINEERING, OSMANIA UNIVERSITY

Table of Contents

Introduction

This report presents a comprehensive study on denoising scanned invoice images, conducted during an internship. Scanned invoice images often suffer from noise due to various factors such as scanning artifacts, paper texture, and environmental conditions, making denoising a crucial preprocessing step for improving readability and ensuring accurate data extraction. The study explores and implements a variety of denoising techniques, including traditional methods (Gaussian filter, Median filter), advanced methods (Non-Local Means, Wavelet Transform), and machine learning approaches (Convolutional Neural Networks, Autoencoders).

The dataset used for this study consists of scanned invoice images with varying noise levels. Each denoising method was evaluated using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics along with text extraction using OCR tools used to assess performance. The results indicate that while traditional methods provide basic denoising capabilities, advanced techniques, particularly those based on machine learning, offer superior performance in terms of noise reduction and detail preservation. The Non-Local Means and Anisotropic Diffusion models were identified as the most effective approaches.

The report concludes with a discussion of the strengths and weaknesses of each method, challenges encountered, and suggestions for future research. This study provides valuable insights into the application of different denoising techniques for enhancing the quality of scanned invoice images, with implications for broader document processing workflows.

Methodologies

This section describes each method used, the implementation process, and the outcomes.

Wavelet Denoising with Soft Threshold

Theoretical Background: Wavelet denoising involves decomposing an image into wavelet coefficients and applying a threshold to these coefficients to remove noise. The soft thresholding method shrinks the coefficients towards zero, reducing noise while preserving significant image features.

Implementation: We applied wavelet transform to the scanned invoices, followed by soft thresholding on the wavelet coefficients. The inverse wavelet transform was then used to reconstruct the denoised image.

Results and Observations: The method effectively reduced noise but caused some blurring of the characters. OCR using Pytesseract struggled to accurately extract text due to this blurring, resulting in poor OCR performance.

Wavelet Denoising with Combinations of Thresholds

Theoretical Background: Different combinations of wavelet thresholds were explored to enhance the denoising performance. This approach involves varying the threshold values and applying multiple thresholding techniques.

Implementation: Various thresholding combinations, including universal, SURE (Stein's Unbiased Risk Estimate), and minimax thresholds, were tested. Each combination was applied to the wavelet coefficients before reconstructing the denoised image.

Results and Observations: While some combinations showed marginal improvements, the OCR accuracy did not significantly improve compared to the soft thresholding method.

Autoencoder Approach

Theoretical Background: Autoencoders are a type of neural network used for unsupervised learning. They are designed to learn efficient codings of input data and can be used for denoising by learning to reconstruct clean images from noisy ones.

Implementation: An autoencoder was trained using a dataset of noisy and clean images from Kaggle. The trained model was then applied to the scanned invoices to produce denoised outputs.

Results and Observations: The autoencoder failed to generalize well to the scanned invoices. The denoised images did not improve OCR results, indicating that the model was not able to effectively remove noise in this context.

Bilateral Filtering

Theoretical Background: Bilateral filtering is a non-linear, edge-preserving, and noise-reducing filter. It smooths images while preserving edges by taking into account the intensity differences between pixels.

Implementation: Bilateral filtering was applied to the scanned invoices using different parameter settings for the spatial and intensity domains.

Results and Observations: This method effectively reduced noise while preserving edges. The OCR results showed improvement but were still not optimal.

Anisotropic Diffusion

Theoretical Background: Anisotropic diffusion is a technique used to reduce image noise without removing significant parts of the image content, typically edges, which are important for interpreting images.

Implementation: The anisotropic diffusion process was applied, followed by Gaussian blur and non-local means thresholding to further enhance denoising.

Results and Observations: This combination significantly improved the denoising quality. The OCR results were better compared to previous methods, with more accurate character extraction.

NSDCT Transform

Theoretical Background: The NSDCT (Non-Subsampled Discrete Cosine Transform) is a multi-resolution transform used for image denoising. It provides a redundant representation, improving denoising performance.

Implementation: The NSDCT transform was applied to the scanned invoices. The denoised images were reconstructed from the transformed coefficients.

Results and Observations: Despite its theoretical advantages, the NSDCT transform did not perform as expected. The OCR results were not satisfactory, indicating inadequate noise reduction.

Dictionary Learning

Theoretical Background: Dictionary learning involves finding a dictionary (a set of basis functions) that allows sparse representations of data. It is used for denoising by learning a sparse representation of the image.

Implementation: A dictionary was trained using a set of clean and noisy images. The denoising was performed by reconstructing the images using sparse coding with the learned dictionary.

Results and Observations: This method was computationally expensive and did not yield efficient denoising results. The increased computation time did not justify the marginal improvement in denoising quality.

Various Threshold Techniques

Theoretical Background: Thresholding techniques, including binary, binary inverse, to zero, to zero inverse, adaptive mean, and adaptive Gaussian, are used to convert grayscale images to binary images based on different criteria.

Implementation: Each thresholding technique was applied to the scanned invoices. The effectiveness of these techniques was evaluated both individually and in combination with other denoising methods.

Results and Observations: Binary thresholding performed better comparatively. These techniques were more effective when used in combination with other algorithms, enhancing the overall denoising performance.

OCR Testing with Pytesseract

Theoretical Background: OCR (Optical Character Recognition) is the process of converting different types of documents, such as scanned paper documents, PDFs, or images, into editable and searchable data.

Implementation: Pytesseract, an OCR tool, was used to test the effectiveness of each denoising method. The accuracy of character extraction was the primary metric for evaluating OCR performance.

Results and Observations: The OCR results varied across different denoising methods. The combination of anisotropic diffusion with Gaussian blur and non-local means thresholding yielded the best OCR accuracy.

PCA and Non-Local Means

Theoretical Background: Principal Component Analysis (PCA) is a statistical technique used to emphasize variation and capture strong patterns in a dataset. Non-local means is a denoising algorithm that averages pixels with similar patches.

Implementation: PCA was applied to reduce dimensionality and enhance important features, followed by non-local means and Gaussian blur for denoising.

Results and Observations: This combination effectively reduced noise while preserving essential features of the documents. The OCR results showed significant improvement, making this method one of the most effective in the study.

ICA Method

Theoretical Background: Independent Component Analysis (ICA) is a computational method for separating a multivariate signal into additive, independent components. It is commonly used in signal processing and data analysis.

Implementation: ICA was applied to the scanned invoices for denoising. Additional methods, such as non-local means and bilateral filtering, were combined with ICA to enhance performance.

Results and Observations: ICA did not work well for denoising in this context. Even with additional filtering techniques, the denoised images did not improve OCR accuracy significantly.

Fractal-Based Methods, Wiener Filter, and Total Variation

Theoretical Background: Fractal-based methods use fractal theory to remove noise. The Wiener filter is a statistical filter used to minimize the mean square error. Total variation denoising minimizes the total variation of the image, preserving edges.

Implementation: Each of these methods was applied to the scanned invoices. Their performance was evaluated based on visual inspection and OCR results.

Results and Observations: These methods did not yield the expected results. The denoising was inadequate for improving OCR accuracy, indicating that these approaches were less effective for this particular application.

Performance Metrics

Theoretical Background: PSNR, SSIM, and MSE are common metrics used to evaluate the quality of denoised images. PSNR measures the peak signal-to-noise ratio, SSIM assesses structural similarity, and MSE calculates the mean squared error between the original and denoised images.

Implementation: These metrics were computed for each denoising method to quantitatively assess their performance. Higher PSNR and SSIM values, along with lower MSE, indicate better denoising quality.

Results and Observations: These metrics provided a useful benchmark for comparing different methods, helping identify the most effective denoising techniques.

Streamlit Application with Final Approaches

Overview To facilitate user interaction with the denoising techniques developed during this internship, a Streamlit application was created. This web application allows users to upload scanned invoices and obtain denoised results using three different approaches: PCA with Non-Local Means, Anisotropic Diffusion with Gaussian Blur, and Bilateral Filtering. Additionally, the application includes a feature to perform OCR on denoised documents, enabling users to extract text and choose the best denoising method based on OCR accuracy.

User Interface and Features

Document Upload The application starts with a simple and intuitive interface where users can upload their scanned invoice documents. The document upload section is straightforward, allowing users to select and upload their files easily.

Denoising Methods Once the document is uploaded, the application processes the document using three different denoising methods. The results are displayed side-by-side for easy comparison, allowing users to visually assess the effectiveness of each method.

1. **PCA with Non-Local Means and Gaussian Blur**
    - This method combines Principal Component Analysis (PCA) with Non-Local Means and Gaussian Blur to enhance important features while reducing noise. The denoised image is displayed for the user to view.
2. **Anisotropic Diffusion with Gaussian Blur**
    - Anisotropic Diffusion is applied along with Gaussian Blur to preserve significant parts of the image content, typically edges, while effectively reducing noise. The resulting denoised image is shown to the user.
3. **Bilateral Filtering**

- Bilateral Filtering is used to smooth the image while preserving edges. This method considers both spatial and intensity differences between pixels to perform edge-preserving denoising. The denoised image is displayed for user assessment.

OCR Extraction At the bottom of the application, there is a button labeled "OCR". When clicked, this button performs Optical Character Recognition (OCR) on each of the denoised documents using Pytesseract. The extracted text from each denoised document is displayed, allowing users to compare the OCR results and determine which denoising method produced the best output.

Resources

Dataset1: https://www.kaggle.com/datasets/pdavpoojan/the-rvlcdip-dataset-test

Dataset2: https://www.kaggle.com/c/denoising-dirty-documents

https://www.techscience.com/cmc/v71n1/45466

https://towardsdatascience.com/denoising-noisy-documents-6807c34730c4

Challenges

Computational Limitations One of the significant challenges faced during the internship was the limitation in computational power. This constraint prevented extensive exploration of deep learning methods, which often require substantial computational resources. As a result, the focus was shifted to rule-based methods, which are less computationally intensive but may not always provide the best results.

Diverse Invoice Formats The variety of invoice formats in the dataset posed a challenge for developing a one-size-fits-all denoising solution. Different invoices had varying levels of noise, text styles, and layouts, making it difficult to design a method that performed well across all samples.

OCR Accuracy Achieving high OCR accuracy was a persistent challenge. While some denoising methods effectively reduced noise, they also introduced artifacts that degraded OCR performance. Balancing noise reduction with the preservation of text clarity was a critical aspect of the research.

Conclusion

Throughout this internship, various denoising methods were explored, implemented, and tested. While some methods showed promise, others did not perform as expected. The combination of anisotropic diffusion with Gaussian blur and non-local means thresholding, as well as PCA with non-local means, yielded better results compared to other methods. The findings highlight the importance of combining different techniques for effective denoising and improved OCR accuracy. The Streamlit application further illustrates the practical implementation and effectiveness of these methods, providing a valuable resource for users needing to denoise and extract text from scanned documents.