A Composite Technique for Degraded Document Image Binarization

ITRI617

BSc. Hons. Computer Science and Information Systems Prof G. Drevin

Computer Science and Information Systems

Potchefstroom Campus NWU

Johan Venter

October 29, 2022

Contents

Chapter 1

Abstract

1.1 English

Before the digital age, all information was stored as written documents, but in a world overflowing with powerful processing tools, there emanates a need to digitise these documents for analysis, inference, data extraction etc. in a meaningful way. Stored for long periods, historical documents in archives tend to degrade. In this context, document degradation is defined as any artefact on the document that can cause ambiguity in differentiating the content of the document from its background. Therefore, a meaningful solution for a digital copy of a written document would contain only the useful content in the image, rid of the degradation. The process of extracting the textual content from the background and degradation is known as document image binarization [SLT12]. The purpose of this project was to secure an understanding of the current state of document image binarization, the techniques and methods used, the obstacles encountered, the efficacy of current methods and finally produce an artefact that demonstrates a solution using information gained from research. The final version of the artefact exists as a composite implementation of existing ideas and methods that were chosen by implementing and comparing various image processing techniques. The program receives a degraded document image as input and transforms it into a binary image.

1.2 Afrikaans

Voor die digitale era is alle inligting as geskrewe dokumente gestoor, maar in 'n wêreld wat oorloop van kragtige verwerkingsinstrumente, ontstaan daar 'n behoefte om hierdie dokumente vir analise, afleiding, data-onttrekking ens op 'n sinvolle wyse te digitaliseer. Geskiedkundige dokumente in argiewe wat vir lang tydperke gestoor word, is geneig om te verneder. In hierdie konteks word dokumentdegradasie gedefinieer as enige artefak op die dokument wat verwarring kan veroorsaak om die inhoud van die dokument van sy agtergrond te onderskei. Daarom sal 'n sinvolle oplossing vir 'n digitale kopie van 'n geskrewe dokument slegs die nuttige inhoud in die beeld bevat, ontslae van die agteruitgang. Die proses om die teksinhoud uit die agtergrond en agteruitgang te onttrek staan bekend as dokumentbeeldbinarisering [SLT12]. Die doel van hierdie projek was om 'n begrip te kry van die huidige stand van dokumentbeeldbinarisering, die tegnieke en metodes wat gebruik word, die struikelblokke wat teëgekom word, die doeltreffendheid van huidige metodes en uiteindelik 'n artefak te produseer wat 'n oplossing demonstreer deur gebruik te maak van inligting verkry uit navorsing. Die finale weergawe van die artefak bestaan as 'n saamgestelde implementering van bestaande idees en metodes wat gekies is deur verskeie beeldverwerkingstegnieke te implementeer en te vergelyk. Die program ontvang 'n gedegradeerde dokumentbeeld as invoer en omskep dit in 'n binêre beeld.

Chapter 2

Introduction

2.1 Project Description

Document Image Binarization is a field of study concerned with the phase of image pre-processing where a document image is transformed into a bi-level image by separating the text from its background [SLT12]. This step precedes the transition to higher-level processing endeavours such as OCR.

2.2 Problem Description and Background

2.2.1 Degradation

There are many obstacles encountered in the process of binarizing historical document images. The main issues are centred around the degradation of written documents. Degradation in this context is seen as a loss of quality that limits the performance of image processing systems [Bai07]. In the case of substantially degraded document images, binarization becomes an increasingly difficult task, since the degree and nature of the degradation can be irregular and unpredictable, differing substantially from image to image. Handling document degradation is the most difficult part of image binarization in that there are many categories of degradation that vary substantially in their characteristic qualities. This property makes the development of a model for degradation extremely difficult. Moreover, each image is usually comprised of numerous forms of degradation in different locations with no regular or predictable pattern, and this forms the crux of our problem.

Natural Environment

The key issue here is the physical deterioration of the original document. Common traits include blur, varying contrast, smudges and blotches, bleed-through text, variable character intensities and stroke widths, artefacts that cover entire sections of text, non-uniform noise or illumination etc. [GPP06] [Ait+22]. Although the degradation of a typical document image consists predominantly of the deterioration of the original physical document, each subsequent step in the processing of an image introduces some form of noise and artefacts.

Scanning Induced

A scanning device is used to obtain a digital version of an image. Due to unavoidable processes, a whole new range of noises and artefacts are introduced to the digital document. Some corrections are automatically made by the scanning devices, mostly colour corrections such as "gamma correction" which adapts the scanned image to be displayed accurately on a monitor [Smo11], but other noise patterns persist, such as noise due to the inevitable material properties of the light sensors used in scanning devices and the way photons are detected or the small imperfections in the sensor's movements

and architecture [Smo11]. Examples include the scanning of an image that produces reflection artefacts or discrepancies due to an uncalibrated sensor. The quality of the sensor will also affect the resolution of the digital result.

After an image has been scanned, it is compressed to an image format like JPEG. Since this is a lossy compression standard, a lot of the original image data is discarded or modified along with other artefacts that are introduced to store the image in a compact format [Esk16]. These changes may be undetectable to the human eye, but is still a loss of original data and will affect the binarization process [Bai07].

2.2.2 Methods and Approaches

Due to the random nature of the image degradations, the development of techniques that accurately model degradations on a wide range of images is met with great difficulty. As such many different approaches have been made towards the advancement of the field and have popularized Image Binarization as a research area [Ait+22] with international competitions such as the recurring DIBCO international competition dedicated to advancing the field.

When developing a method that binarizes degraded document images effectively, prior analysis of the document's features plays the most significant role. That requires analysis of the characteristics of the relevant document images and how they differ, as well as how the document content and properties affect the processing method. This leads to a better understanding of the typology and characteristics of these digital documents and how to enhance the desired properties or get rid of unwanted properties. Papers on existing methods proceed on this assumption directly or indirectly. Since our concern is with the binarization of exclusively textual documents and their representation as bi-level images, we can disregard multi-spectral image analysis and focus on monochromatic/grey-level images.

Global, Local and Adaptive Methods

From this point, research papers on this topic typically approach the enhancement method of document images using a step-by-step scheme, where each step is a process based on some existing method developed within the context of general image processing that modifies or extracts information from the document image. These methods broadly fall into global or local categories based on how thresholds and enhancements are calculated. Local methods are more adaptive since they calculate statistics of pixels in a neighbourhood around a pixel [GPP06]. Global methods provide high-level information on the overall characteristics of the image and are less adaptive since it calculates a single parameter using statistics about the entire image that is then used on every pixel [GPP06]. Adaptive methods embrace the benefits of both categories by modifying the image using a combination of global and local statistics. This hybrid approach produces better results on a wider range of degraded documents.

Scanner Models

Text Documents are scanned in some way to produce a digital version. Different types of noise are introduced in the scanning process. As indicated by [Smo11], the most common noise patterns have been modelled successfully, but as [Esk16] points out, there are many more unmodelled noise profiles present in scanned document images that aren't modelled as easily and are mostly modelled empirically.

Denoising

Noise is random high-frequency variations in the colour intensities of an image that look like minor disturbances in the context of the entire image. The high-frequency nature of image noise is what denoising methods rely on when attempting to remove it. There is a myriad of different denoising techniques broadly falling into two main categories, spatial filtering and transform domain filtering and each has numerous sub-categories [Mot+04]. Some denoising filters are modelled to remove specific types of noise such as Gaussian noise, Poisson noise, salt and pepper noise etc.

Contrast Analysis

Contrast filters identify the areas in an image with high colour intensity disparity or high contrast between neighbouring pixels. This is useful in locating the edges of character strokes. The advantage of using a local contrast or gradient filter such as the one used by [SLT12] is that the image intensity levels become normalized, meaning that previously darker or lighter areas in the image now have more similar intensities while retaining the same relative contrast as before. This is a very powerful result since a global thresholding filter can now reliably separate the text and the background without discarding information in previously lighter or darker areas of the document image.

Thresholding

Image Binarization is confronted with many obstacles and has some limitations at the moment, such as the need for a robust thresholding method which is an unsolved problem [SLT12]. Thresholding is the common process between each binarization technique. Threshold filters transform an image into a bi-level image, that is, an image with pixels taking on only one of 2 distinct intensity values. This is done by comparing each pixel intensity to a calculated threshold value and setting the new intensity to 0 or 1 depending on the comparison result. Thresholding can also fall into local, global or adaptive categories. Thresholding typically marks the final step before the image enters the post-processing stages.

Learning Models

Learning models are used in conjunction with all the previously mentioned categories to aid in parameter estimation and higher-level image modifications. Advanced learning and clustering models are very popular today and are used for denoising, thresholding and character recognition even before the image enters the post-processing stage.

Post-Processing

Up to this point, the image has been modified in such a way as to simplify post-processing procedures. Since, at this point, the image is in a bi-level format, higher-level, more complex processing methods can now be used such as Edge and Stroke detection, Stroke width estimates for identifying remaining noise, learning models and OCR.

Reseach

This section has outlined the current approaches and challenges related to the field of document image binarization. The scope of this study will entail acquiring a grounded understanding of the specific techniques implemented and their inner workings. Their rationales will be evaluated and compared with others to identify candidate methods. The candidate methods will be implemented and evaluated for the procurement of their respective benefits and drawbacks. This study will aim to deliver a program that is composed of various implementations of the best-performing methods.

2.3 Aims and Objectives of the Project

2.3.1 Aims

This study aims to research then collect, develop and adapt a collection of methods and models that can identify and isolate general characteristics of degraded document images. The second goal is to then adapt an existing generic method by integrating these models, which takes a digital document image as input and produces a binarized image that can be used in post-processing endeavours.

2.3.2 Objectives

- Aquire an understanding of existing techniques used in image processing and evaluate their rationales.
- Compile a collection of candidate techniques and mathematical and statistical models that can best isolate and remove the unwanted artefacts in a degraded document image.
- Compile and test different adaptations of the paper by [SLT12] that make use of the previously selected models to arrive at a model.
- Implement and optimize the developed model.
- Deliver an implementation of the method that successfully separates the textual content from the background of any given input image.

2.4 Procedures and Methods

2.4.1 Paradigmatic Perspective

Logical Positivism

Positivism as a research paradigm is grounded on a proposition about the nature of reality and its properties, and the methods by which information and understanding can be obtained. Positivism departs from the belief that reality is composed of material objects that have properties where statements can be derived about these properties using one's senses [Put12]. This idea rests on actual, objective reality, meaning that objects exist absolutely and independently of the perceiver. Reality can therefore be described in terms of constant universal properties, presentable as perceivable truths. Positivism takes reality as a collection of interdependent objects governed by laws and symmetries that conduct all occurring events. As assumed by determinism, observed events are dependent on other factors and an understanding of the relevant factors allows the prediction and control of events [KK17].

Positivism assumes that everything knowable can be discovered. The paradigm introduces the methods by which these truths can be obtained. Facts can be empirically collected or derived by using the scientific method. The scientific method entails the use of logical reasoning, deduction, the formulation of hypotheses, experimentation and mathematical methods to derive conclusions, therefore, creating an understanding of a part of reality [KK17].

Application

This research paradigm is well suited for this study since it relies on the acquisition of material text documents that have certain characteristics. These characteristics include the textual content of the documents as well as the associated artefacts.

The proposition of objectivity is appropriate in this context since the text on a document consists of characters that can each be designated a ubiquitous category. That is, the method assumes the use of a document with textual content that will be transformed into an image containing the same objective textual content. The product of the method is an artefact delivering a distinct separation between text and background, implying that the ideal solution produces an image containing only the textual content of the original image. The performance of this artefact will consequently be evaluated by empirical methods.

Unwanted artefacts and degradation can therefore be viewed deterministically i.e., there are external factors that produce these artefacts and an understanding of them can be used to shape empirical models. By using scientific methods such as mathematical and statistical analysis, empirical modelling and experimentation one can analyse the documents and model their properties and dependent factors.

Methods

This study will start with research on digitised document images i.e., the associated processes in creating them and their artefacts as well as insight into the properties and characteristics that distinguish them from other types of scanned images. The resulting context will be used to explore existing document image binarization solutions and the challenges they encounter. The existing solutions will then be decomposed into their distinct components, that is, broken down into the individual parts of the process, that can each be described by the specific goal it aims to achieve. These techniques will undergo experimental analysis to compare their results. Resulting inference on these results will be used in the development of an artefact that will demonstrate an implementation of the chosen methods.

2.5 Project Management and Plan

2.5.1 Plan

2.5.2 Scope

The research of this project will be confined to the binarization of textual document images only. The textual content of the documents can be either typed or written text. This study will focus on a wide range of degraded images. The resulting artefact will aim to take a degraded document image as input and deliver a modified bi-level image rid of all degradations yet containing the same text as the original image. Since the artefact will produce a bi-level image, research on the nuances regarding colour and multi-spectral processing and analysis will be omitted. Using the paper by [SLT12], the study will compose an adaptation of their technique as an artefact.

2.5.3 Limitations

The topic of document image binarization is a well-established field with numerous years of relevant research. This means that due to time constraints and research background, this study cannot be comprehensive at that scale. Due to the nature of the degradation of the documents as well as other previously mentioned factors, there are some trade-offs involved in removing these defects that result in either a loss of important information or the persistence of prominent degradations. This study is primarily concerned with existing techniques with a secondary objective of novel development in the field.

2.5.4 Risks

As of the time when this was written, no risks were identified in conducting this study or the development of the artefact since it makes use of open-source libraries and datasets and does not collect any personal information from anyone. The results of this project are also risk-free since it dabbles in a well-researched area.

2.6 Development Platform, Resources and Environments

The chosen models and methods will be implemented as an independent script within the python programming language since it is a language optimised for prototyping and it is a simple language but most importantly, it provides substantial support, environments and libraries for image processing. The open source python libraries that will be used are:

- numpy [Har+20]
- scikit-image [Wal+14]
- scipy [Vir+20]

A simple front-end application will be constructed using the Angular Web Framework [22a] for displaying the input and output images and navigating between them for demonstration as well as development and testing. The front-end application will retrieve the results from the python script through a locally hosted Nodejs [Nod22] server that will execute the package on the chosen image since it is a simple reliable way of serving images and results. The implementation of the models in python will not depend on any of the auxiliary architectures.

The images chosen for testing and demonstrations are provided by DIBCO 2016 Handwritten DocumentDataset. They were chosen since they are open-source, and therefore free from legal implications (within regulation). These datasets are also the ones used by the yearly DIBCO competition.

2.7 Ethical and Legal Implications and Dealing with These

see addendum A

Chapter 3

Literature Review

This chapter will discuss the ideas and challenges surrounding the binarization of degraded document images. After discussing the image processing pipeline, the first subsection discusses what degradation is, what its causes are, how modelling is approached and lastly presents some current models.

3.1 Image Processing Pipeline

Image binarization falls into the pre-processing stage where degradation and defects are compensated for. The pre-processing stage forms a pipeline of processes that each modify its input before passing it to the next process in the pipeline. The first of these is called the pre-processing stage where degradation and defects are compensated for [Ram+05]. Since this project is concerned with document images, the binarization of the document also falls into the pre-processing stage.

Since document image binarization aims to separate the textual content from the background by removing degradations, the first question arises: What does degradation on an image look like and how does one model it?

3.2 Document Image Degradation

Degradations or defects on a document image refer to any properties of actual document images that deviate from what can be considered the ideal image that reduces the efficiency of image processing systems [Bai07]. When paper documents are printed, copied, faxed and scanned, the documents degrade. Even if it seems insignificant to the human eye, this quality loss can reduce the accuracy of even the most cutting-edge text recognition systems (OCR). Additionally, there is mounting evidence that the quantity and representativeness of training sets as well as the feature selection have a considerable impact on the accuracy of stubborn picture pattern recognition tasks [Bai07]. Therefore, the performance of these algorithms hinges on the quality of the binarization result.

3.2.1 Operational Degradation

This type of degradation involves the defects produced by the equipment used in obtaining a digital copy of a given document image. An image is converted to digital form using a scanning device. Unavoidable procedures result in the introduction of a wide variety of noise and artefacts into the digital document. Other noise patterns persist, such as the small imperfections in the sensor's movements and architecture. Some corrections are automatically made by the scanning devices. These corrections are mostly colour corrections, such as "gamma correction," which adapts the scanned image to be displayed accurately on a monitor [Smo11]. Examples include the scanning of an image that produces reflection artefacts or discrepancies due to an uncalibrated sensor. The quality of the sensor will also affect the resolution of the digital result. Since this type of degradation is caused by the scanner, it tends to be easier to model

since it is independent of the image and by analysing the effects of the scanner on different images, one can model the defects of the scanner.

An image is compressed into an image format like JPEG after it has been scanned. Since this is a lossy compression standard, a significant portion of the original image data is lost or altered [Esk16].

3.2.2 External Degradation

The physical deterioration of the original document is the main problem in this situation. Examples of this type of degradation include blur, fluctuating contrast, smudges and blotches, bleed-through text, variable character intensities and stroke widths, artefacts that cover entire areas of text, non-uniform noise or illumination, etc. [GPP06] [Ait+22]. This type of degradation refers to the defects related to the physical state of the document, independent from the result produced by the apparatus used to scan the document. This category of degradation can be the most difficult to model as it cannot be generalized since different degradations can appear at different intensities at various locations in the document image. This can be further illustrated by the fact that these types of degradation sometimes overlap with what is not considered degradation. As an example, a model that may be good at identifying and removing ink-smearing may compromise the quality of the document in areas where ink-smearing is not present.

3.2.3 Inevitable Disparities

This type of degradation is described by the inevitable physical properties of light, reflection, and the way that scanning devices scan images as well as properties like the resolution of the images. Some of these degradations must be modelled and analysed statistically [Bai07], but for some defects, there are no existing models and are considered roadblocks in the image pre-processing pipeline.

3.3 Issues With Modelling

[Bai07] notes that the main components affecting the modelling process are:

- **Parameterization**: The observed degradation must be able to be described by a fixed set of numerical parameters. Otherwise, the model is of no use.
- **Randomization**: If the degradation is modelled as having properties that behave in a probabilistic manner, their distributions should be parameterized in the model.
- **Validation**: Since some of the model parameters will inevitably be probabilistic, one needs to compensate for this by accounting for certain margins of error.
- **Parameter Estimation**: Once the model is developed, a distribution fitting the real distribution can be generated by tuning the parameters.
- **Correlation**: When modelling using statistics and probability distributions, one can easily mistake correlation for causation, therefore the model must be thoroughly evaluated by predicting the output of changes in parameters and verifying those predictions. Unexpected results may indicate that the model is incomplete.

3.4 Modelling

Document image degradation can be modelled in two main ways. The first is to analyse and model the physical properties of the factors that cause these defects. This can in turn deliver a degradation model based on the workings of the environment that the images are subjected to. Although this method will theoretically yield the most accurate, reliable results, this approach can quickly become

overcomplicated as [Bai07] points out. The second approach is to model the degradation empirically. This involves developing a model that can replicate the effects of the degradation, disregarding the cause. The main way this is done is by using statistical analysis. Of course, these methods can be combined to form a hybrid model.

3.4.1 Noise

Image noise, which is typically introduced by electrical noise, is the random variation in brightness or colour information in images. It can be created using a scanner or a digital camera's image sensor and circuitry. This unwanted by-product of image capturing known as image noise obscures desired information. The devices used to scan documents introduce various forms of noise to the document. Due to the random nature of noise, it is typically modelled statistically. Since this topic has been the subject of research for a significant amount of time (more than 80 years), existing noise models are very sophisticated and although most models are statistics based, the papers also consider the underlying physical structure of the scanning device in their models such as the scanner model depicted by [GSW07] or the range of noise patterns by imaging sensors modelled by [LFG06].

The following subsections describe common statistical methods used for noise modelling.

Gaussian Noise

The most common type of noise found in images is Gaussian noise. It is an additive noise model. The probability density function p of a Gaussian random variable z is given by

$$p_G(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

where z is the pixel intensity, μ the mean pixel intensity and σ its standard deviation.

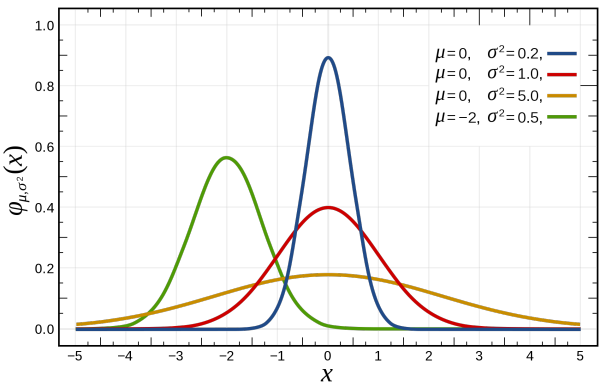

Figure 3.1: Gaussian PDF with different parameters [Con22a]

Impulse Noise (Salt and Pepper Noise)

Impulse noise is characterised by sharp and rapid changes in the image signal. It appears as sparsely distributed white and black pixels.

Shot Noise

Shot noise or Poisson noise occurs due to the statistical character of electromagnetic waves and the random fluctuation of photons [BJ15] and is modelled by the Poisson distribution. The Poisson distribution is a function of a discrete random variable X.

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

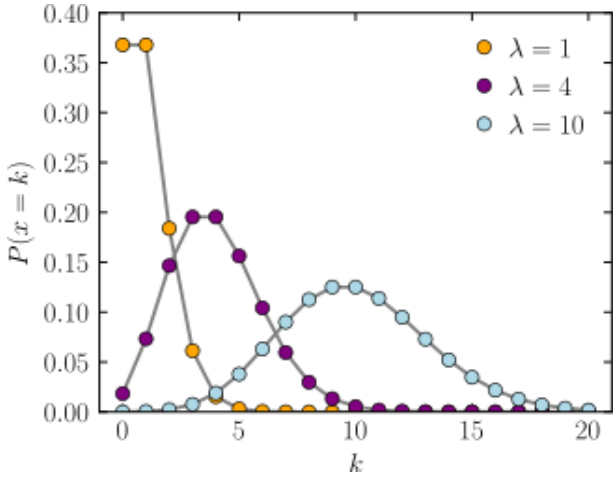

Figure 3.2: Poisson PDF with different parameters [Con22b]

with $\lambda > 0$

Other noise models include Rayleigh noise, Gamma noise, Exponential noise and Uniform noise each having a probability density function that describes the distribution of the noise in an image [Gon+].

Noise can be classified according to which domain it is filtered in. Noise removal can either be done in the spatial domain or the frequency domain. Images must undergo the Fourier Transform to map them onto the frequency domain. In spatial domain filtering there exist some popular methods such as Mean, Wiener and median filtering. Filtering in the frequency domain yields many more opportunities such as filtering in the wavelet domain using linear, non-linear, and other models.

3.4.2 Scanner Degradation Model

An image is scanned by a scanning device using a light-sensitive sensor. This sensor senses the light reflected by the document in some neighbourhood of a pixel in the image close to the sensor. Since the surrounding areas affect the signal received by the sensor, some distortion is introduced that can be modelled. This type of distortion is typically modelled as a Point Spread Function (PSF). [Smi98] provides a method for estimating a point spread parameter δ_c that is dependent on the period of a detected pulse τ, the width of a square pulse ω and some pre-decided thresholding value Θ, delivering a piecewise function:

$$\delta(\tau) = \begin{cases} \frac{\tau}{2} & A \\ \omega(\frac{1}{2} - \Theta) & B \\ -\frac{\tau}{2} & C \end{cases}$$

A, B and C represent sections of the scanned image with decreasing density (number of black pixel sections and white pixel sections per area). This model is used to generate synthetic characters that mimic the point spread in the signal received by a scanner and is implemented in the training models of classifier systems as well as evaluation benchmarks for OCR.

Θ_{grey}	Scanned Character	Synthetic Characters	Scanned Character	Synthetic Characters	Scanned Character	Synthetic Characters
90	c	ccc	e	eee	m	mmm
150	c	ccc	e	eee	m	mmm
220	c	ccc	e	eee	m	mmm

Figure 3.3: Synthetic Characters [Smi98]

3.4.3 Character Degradation Modelling

The most common type of degradation is some sort of distortion of the characters of the document. These common types include ink blotches and smears. This type of degradation is characterised by dark areas over or around a character. The paper by [Kie+12] proposes a method that models this as a region of noise, They use a formula that describes a region of noise as an elliptic area wherein a random generation function following the normal distribution $N(\mu, \sigma x^2)$ is used to generate noise where σ is an input parameter and μ is calculated based on pixel values in the segment where the noise will be added.

The results of their model by varying the σ parameter:

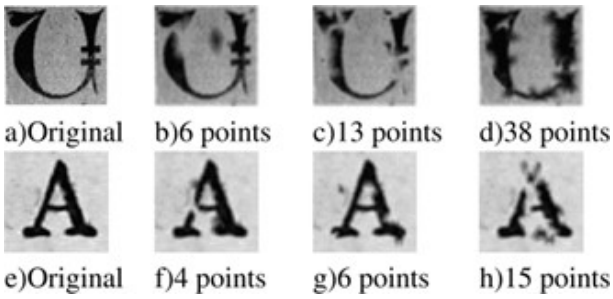

Figure 3.4: Varying the σ parameter [Kie+12]

This model sufficiently mimics the defects present in documents affected by natural elements. In the paper, they note that the performance of OCR goes down as the number of defects increases, and as such these artificial characters are also used for training classifier systems.

3.5 Implementation

3.5.1 Noise Removal

The first step in document binarization is to remove the various forms of noise present in the images. The method used is dependent on the characteristics of the noise identified as using the wrong model will not remove the noise effectively and compromise the quality of the image.

Median Filter

The median filter traverses the pixels of the image and replaces the values of the pixels with the median value in some neighbourhood around the pixel called a window. The grey values inside the window are ordered and the centre pixel's value is replaced with the median grey value. This technique is nonlinear and highly effective at removing salt-and-pepper noise since sudden jumps in grey value magnitudes are removed. The median filter is however not very computationally efficient since sorting takes a lot of time.

Wiener Filter

Wiener filtering is a popular statistical noise removal method based on minimising the mean square error for stationary Gaussian process [Jin+03]. By this measure, the wiener filter is the optimal filter. The basic model is directed towards finding an estimator $\hat{x}$ for the original image S that minimises:

$$MSE(\hat{x}) = \frac{1}{N} \sum_{i,j \in S}^{N} (\hat{x}(i,j) - x(i,j))^2$$

They can be treated as stationary since Gaussian noise is considered globally independent noise. The filter takes two parameters μ being the signal mean and σ^2 the signal variance. While the global version uses global statistics of the image to minimise the error, the local version is more adaptive and calculates the statistics in different neighbourhoods in the image.

[GPP06] describes the use of an adaptive wiener filter that makes use of the local properties in an image to reduce noise using the following formula:

$$I(x,y) = \mu + \frac{(\sigma^2 - v^2)(I_{source} - \mu)}{\sigma^2}$$

where μ is the local mean, σ^2 the variance in a 3 x 3 window and v^2 is the average variance.

The Wiener-Hunt filter is a variation on the original Wiener filter, that transforms the image into the frequency domain first by using the Fourier transform

$$\hat{x} = F^{\dagger}(|\Lambda_H|^2 + \lambda|\Lambda_D|^2)\Lambda_H^{\dagger}Fy$$

with F and $F^{\dagger}$ the Fourier and inverse Fourier transforms respectively, Λ_H the Fourier transform of the transfer function and λ a damping constant as described by [Wal+14].

Wavelet Filter

The wavelet transform decomposes the image into a collection of wavelets. A wavelet is a wave-like function, that has a finite 'energy' and a symmetric area around the x-axis. The transform can be interpreted as the convolution of a set of wavelets over the image that will output a signal proportional to the similarity between the wavelets and the image.

$$\int_{-\infty}^{\infty} f(x,y) \cdot g(x,y)\, dx$$

Kuan Filter

By substituting the core pixel with a weighted average of the centre pixel and the mean of the values in a square kernel surrounding the pixel, the Radar Kuan filter determines the value for each pixel. Edge pixel values are duplicated to create enough data to filter pixels close to the image's edges.

The main purpose of this filter is to reduce speckle noise. While reducing the loss of radiometric and textural information, it smooths image data without eliminating edges or sharp features in the images. The multiplicative noise model is initially changed by the Kuan filter into a signal-dependent additive noise model. The model is then subjected to the least mean square error criteria [22b]. The smoothed pixel's calculated grey-level value (R) is:

$$R = xW + \mu(1 - W)$$

where

- x is the centre pixel in a neighbourhood
- σ= standard deviation of pixel intensities in the window
- μ= mean of the pixel intensities in the window
- $i = \frac{\sigma}{\mu}$
- $u = \sqrt{\frac{1}{numberofiterations}}$
- $W = \frac{1-\frac{u^2}{i^2}}{1+u^2}$

[22b]

Other Denoising methods were considered such as total-variation denoising and bilateral denoising.

3.5.2 Edge Detection

Edge detection is a technique in image processing that highlights the boundaries of objects in images by identifying areas with high contrast.

Variance Edge Detection

By calculating the variance of the intensities of pixels in a neighbourhood, one can construct a simple basic edge detector:

$$S^2 = \frac{\sum^N (x_i - \bar{x})^2}{N - 1}$$

with N being a dimension of the image

Canny Edge Detection

The Canny edge detection algorithm works as follows:

1. A Gaussian filter is convolved to remove noise.
2. The intensity gradient of the image is calculated by using directional edge operators such as Roberts, Prewitt or Sobel.
3. The gradient magnitude is calculated and used as an upper and lower threshold
4. Track edges by hysteresis.

[Ron+14] outlines the principal part of the algorithm where the edges are identified by calculating the gradient between pixels. The following first-order partial derivative estimations are used on the original image I for pixels $i, j \in I$:

$$E_x(i,j) = \frac{1}{2}(I(i+1,j) - I(i,j) + I(i+1,j+1) - I(i,j+1))$$
$$E_y(i,j) = \frac{1}{2}(I(i,j+1) - I(i,j) + I(i+1,j+1) - I(i+1,j))$$

which can also be written in matrix form as

$$\mathbf{G}_x = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \mathbf{G}_y = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$$

The magnitude is calculated as

$$|\mathbf{G}| = \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2}$$

and the edge direction as

$$\Theta = atan2(\mathbf{G}_y, \mathbf{G}_x)$$

[Ron+14]identifies some problems with the traditional Canny algorithm. The first is that since the convolution window is a 2x2 matrix, it is sensitive to noise. Seeing that the algorithm uses the partial derivatives in the x and y direction, it is not sensitive to edges oriented 90° relative to the major axes.

Due to this, some augmented methods have been constructed that attempt to solve these issues such as those proposed by [Ron+14] and [XH17]

Contrast Edge Detection

An image is constructed using a formula for identifying local image contrast as proposed by [SLT12].

$$C(i,j) = \frac{I_{max}(i,j) - I_{min}(i,j)}{I_{max}(i,j) + I_{min}(i,j) + \epsilon} \tag{3.1}$$

where ϵ is a positive infinitesimal. The denominator scales the input according to the local range of values, thereby normalizing the contrast across the image [SLT12]. Next, a constant a is calculated:

$$a = (\frac{\sigma}{128})^{\gamma},\ \gamma \geq 0 \tag{3.2}$$

σ is the standard deviation of the entire document image intensities. By combining 3.1 and 3.2 we derive the final equation they use to construct the Gradient image:

$$C_a(i,j) = a \times C(i,j) + (1-a)(I_{max}(i,j) - I_{min}(i,j)) \tag{3.3}$$

3.5.3 Thresholding

Thresholding is the technique of converting a grayscale image to black and white by turning just the pixels whose values are higher than a specified threshold into white and the other pixels into black [Gur20]. A straightforward method of segmenting images is image thresholding. This is frequently done to segregate "object" or foreground pixels from background pixels by designating pixel intensities one of two values depending on a comparison with a calculated threshold value.

Binary Thresholding

A binary threshold will set all values larger than a calculated threshold value to 1 and all other values to 0.

$$i(x,y) = \begin{cases} 1 & I(x,y) > T \\ 0 & I(x,y) \leq T \end{cases}$$

with threshold value T

Otsu's Method

Otsu's method is an automatic image thresholding approach. The algorithm returns a single intensity threshold, dividing pixels into the foreground and background classes. This threshold is calculated by optimising for inter-class and intra-class variance.

3.5.4 Post Processing

After the image has been binarized it can finally be used in OCR programs and other sorts of post-processing endeavours such as stroke analysis and other typographic analysis meant to discover patterns and tendencies in written text such as handwriting recognition and adaptive natural typeface generation.

3.6 Conclusion and Closing Remarks

This section has discussed a collection of popular techniques and methods used in binarizing document images. A central idea is identified that gets expanded upon in the different papers. All the techniques and models are based on the identification of some pattern following the development of a function that implements the model. This project will be testing and evaluating some of the discussed methods as well as new ones should their relevance become apparent. There are some existing libraries for use in the python programming language that have an extensive collection of image processing-related functions. Should any of the necessary functions be missing they will be implemented using their theoretical models. After implementing these methods, their performance will be evaluated and compared using the DIBCO 2016 document image dataset. This dataset is comprised of 10 different document images along with 10 ideal images that have no degradation.

Chapter 4

Description of the Artefact

The development of this artefact was inspired by the method used by [SLT12] for document image binarization. Document Image Binarization, in this context, is a field of study within the phase of image pre-processing where a digital version of a written document is transformed into a binary image, separating the text from the noise, degradation and background [SLT12]. Therefore, this artefact had a clear objective, namely, to obtain a digital written document, process the image by removing unwanted artefacts and deliver a binary image containing only the text from the original image.

This artefact, therefore, is a program written in python that receives a document image as input and then passes the image through a pipeline of distinct processes, each modifying the image or extracting information about the image for use in the following process. The resulting program uses three main libraries namely numpy [Har+20], scikit-image [Wal+14] and scipy [Vir+20] in conjunction with custom-developed methods to compose the processes that make up the pipeline.

A document image is provided as input and is converted to a grayscale image for processing. This is done by averaging the colour channels of the image into a single channel. It is then passed through a series of steps that each modify it in some way. The process is comprised of four main steps. The images used for testing and demonstrations are open-source, provided by DIBCO 2016 Handwritten DocumentDataset.

4.1 Development Lifecycle

The general lifecycle of the development of this project involved the iterative research and testing of multiple existing strategies.

4.1.1 Research

The first step in the lifecycle was the research for this project. The research needed to be directed towards gaining an understanding of the problem, the related fields and key terms and jargon related to the problems encountered.

4.1.2 Existing Solutions

Once a firm understanding of the field was established, existing solutions were researched and evaluated. This also required an understanding of the surrounding technologies and fundamentals of the methods used.

4.1.3 Development

Candidate methods, ideas and technoloies were identified and appropriate open source libraries were leveraged where needed. Several methods were implemented, tested and relinquished.

4.1.4 Iterative Approach

Although the main lifecycle of the development of the project was sequential, each step happened iteratively.

4.1.5 Denoising

Wiener Filter

The Wiener filter is optimal by the mean squared error measure, since it is defined by it. This property along with popular use and consistency of results by [Nat13], [GPP06] makes the wiener filter as well as wavelet filters the obvious choice as candidates for denoising the document images.

[GPP06] describes the use of an adaptive wiener filter that makes use of the local properties in an image to reduce noise using the following formula:

$$I(x,y) = \mu + \frac{(\sigma^2 - v^2)(I_{source} - \mu)}{\sigma^2}$$

where μ is the local mean, σ^2 the variance in a 3 x 3 window and v^2 is the average variance.

The selected algorithm is a variation on the original Wiener filter, called the Wiener-Hunt filter that transforms the image into the freqency domain first by using the fourier transform

$$\hat{x} = F^{\dagger}(|\Lambda_H|^2 + \lambda|\Lambda_D|^2)\Lambda_H^{\dagger}Fy$$

with F and $F^{\dagger}$ the Fourier and inverse Fourier transforms respectively, Λ_H the Fourier transform of the transfer function and λ a damping constant as described by [Wal+14].

Wavelet Filter

The wavelet transform decomposes the image into a collection of wavelets. A wavelet is a wave-like function, that has a finite 'energy' and symmetric area around the x-axis. The transform can be

interpreted as the convolution of a set of wavelets over the image that will output a signal proportional to the similarity between the wavelets and the image.

$$\int_{-\infty}^{\infty} f(x,y) \cdot g(x,y)\, dx$$

Other Denoising methods were considered such as total variation denoising and bilateral denoising.

4.1.6 Thresholding

Bibliography

[22a] 2022. URL: https://angular.io/.

[22b] 2022. URL: https://catalyst.earth/catalyst-system-files/help/concepts/orthoengine_c/Chapter_824.html.

[Ait+22] Fatim Zahra Ait Bella et al. "An innovative document image binarization approach driven by the non-local p-Laplacian". In: *EURASIP Journal on Advances in Signal Processing* 2022.1 (2022), pp. 1–18.

[Bai07] Henry S. Baird. "The State of the Art of Document Image Degradation Modelling". In: *Digital Document Processing: Major Directions and Recent Advances*. Ed. by Bidyut B. Chaudhuri. London: Springer London, 2007, pp. 261–279. ISBN: 978-1-84628-726-8. DOI: 10.1007/978-1-84628-726-8_12. URL: https://doi.org/10.1007/978-1-84628-726-8_12.

[BJ15] Ajay Kumar Boyat and Brijendra Kumar Joshi. "A review paper: noise models in digital image processing". In: *arXiv preprint arXiv:1505.03489* (2015).

[Con22a] Wikipedia Contributors. *Normal distribution*. Oct. 2022. URL: https://en.wikipedia.org/wiki/Normal_distribution#/media/File:Normal_Distribution_PDF.svg.

[Con22b] Wikipedia Contributors. *Poisson distribution*. Oct. 2022. URL: https://en.wikipedia.org/wiki/Poisson_distribution#/media/File:Poisson_pmf.svg.

[Esk16] Sébastien Eskenazi. "On the stability of document analysis algorithms: application to hybrid document hashing technologies". PhD thesis. Université de La Rochelle, 2016.

[Gon+] Rafael Gonzalez et al. *Digital Image Processing Third Edition Pearson International Edition prepared by Pearson Education*. URL: http://sdeuoc.ac.in/sites/default/files/sde_videos/Digital%20Image%20Processing%203rd%20ed.%20-%20R.%20Gonzalez%2C%20R.%20Woods-ilovepdf-compressed.pdf.

[GPP06] Basilios Gatos, Ioannis Pratikakis, and Stavros J Perantonis. "Adaptive degraded document image binarization". In: *Pattern recognition* 39.3 (2006), pp. 317–327.

[GSW07] Hongmei Gou, Ashwin Swaminathan, and Min Wu. "Robust scanner identification based on noise features". In: *Security, steganography, and watermarking of multimedia contents IX*. Vol. 6505. SPIE. 2007, pp. 289–299.

[Gur20] Prathima Guruprasad. "OVERVIEW OF DIFFERENT THRESHOLDING METHODS IN IMAGE PROCESSING". In: June 2020.

[Har+20] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[Jin+03] Fu Jin et al. "Adaptive Wiener filtering of noisy images and image sequences". In: *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*. Vol. 3. IEEE. 2003, pp. III–349.

[Kie+12] VC Kieu et al. "A character degradation model for grayscale ancient document images". In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE. 2012, pp. 685–688.

[KK17] Charles Kivunja and Ahmed Bawa Kuyini. "Understanding and applying research paradigms in educational contexts." In: *International Journal of higher education* 6.5 (2017), pp. 26–41.

[LFG06] Jan Lukas, Jessica Fridrich, and Miroslav Goljan. "Digital camera identification from sensor pattern noise". In: *IEEE Transactions on Information Forensics and Security* 1.2 (2006), pp. 205–214.

[Mot+04] Mukesh C Motwani et al. "Survey of image denoising techniques". In: *Proceedings of GSPX*. Vol. 27. 2004, pp. 27–30.

[Nat13] Asoke Nath. "Image Denoising Algorithms: A Comparative Study of Different Filtration Approaches Used in Image Restoration". In: *2013 International Conference on Communication Systems and Network Technologies*. 2013, pp. 157–163. DOI: 10.1109/CSNT.2013.43.

[Nod22] Node.js. *Node.js*. 2022. URL: https://nodejs.org/en/.

[Put12] Hilary Putnam. *Philosophy in an age of science: Physics, mathematics, and skepticism*. Harvard University Press, 2012.

[Ram+05] Rajeev Ramanath et al. "Color image processing pipeline". In: *IEEE Signal Processing Magazine* 22.1 (2005), pp. 34–43.

[Ron+14] Weibin Rong et al. "An improved CANNY edge detection algorithm". In: *2014 IEEE international conference on mechatronics and automation*. IEEE. 2014, pp. 577–582.

[SLT12] Bolan Su, Shijian Lu, and Chew Lim Tan. "Robust document image binarization technique for degraded document images". In: *IEEE transactions on image processing* 22.4 (2012), pp. 1408–1417.

[Smi98] Elisa H Barney Smith. "Characterization of image degradation caused by scanning". In: *Pattern Recognition Letters* 19.13 (1998), pp. 1191–1197.

[Smo11] Andreea Smoaca. "ID Photograph hashing: a global approach". PhD thesis. Saint-Etienne, 2011.

[Vir+20] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[Wal+14] Stéfan van der Walt et al. "scikit-image: image processing in Python". In: *PeerJ* 2 (June 2014), e453. ISSN: 2167-8359. DOI: 10.7717/peerj.453. URL: https://doi.org/10.7717/peerj.453.

[XH17] Li Xuan and Zhang Hong. "An improved canny edge detection algorithm". In: *2017 8th IEEE international conference on software engineering and service science (ICSESS)*. IEEE. 2017, pp. 275–278.