# Binarization of the Minutes of the Early Synod Meetings of the Reformed Churches in South-Africa

ITRI 671
Zander Labuschagne 23585137
B.Sc.(Hons) Computer Science & Information Systems 2017

Dissertation submitted in partial fulfillment of the requirements

for the degree B.Sc.(Hons) in Computer Science and Information

Systems at the Potchefstroom Campus of the North-West University

Supervisors: Prof. G.R. Drevin & Mev. S. Campher

March 2017

# Contents

## List of Figures

# 1 Introduction

This project will determine the binarization algorithm(s) or combination of binarization algorithms which best preserves the quality of the text of historical handwritten document images, more specifically the document images containing the minutes of the early synod meetings of the Reformed Churches of South-Africa.

The main contribution of this project will result in a detailed description of an algorithm which proved to be the best in terms of preserving the quality of historical handwritten document text, as well as the actual algorithm in Matlab syntax along with the proof supporting the solution's performance whether it be a comparison of binarized images or statistical data. The algorithm will also be implemented and used to binarize the document images mentioned earlier.

# 2 Problem Description

*If I had an hour to solve a problem I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions.*

*Albert Einstein*

## 2.1 Background

The idea for this project came to mind when the Theological Library of the North-West University at the Potchefstroom Campus saw the need for the digitalization of historical documentation. The library realized that if nothing is done to prevent further degradation on their archives it will not last forever even in the carefullest hands. The greater picture is to digitalize all the archives in order to keep the history in tact even after the original documentation perishes completely, this will also in turn provide easier access to many more people who are interested in this information as well as providing easier and quicker search methods. However this project will currently focus on the binarization objective only.

Binarization in image processing is the process of discriminating the foreground from the background in an image by changing the foreground pixels or text to black and the background or paper pixels to white [7]. Binarization is also considered by many computer scientists as a critical step in the processing of images because it has the greatest impact on the quality grade of all the processing

that follows such as document analysis [5], thus a good binarization result would greatly improve the results of page-segmentation, optical character recognition and any other subsequent processing [1], while a bad binarization result will cause any following processing techniques to yield inadequate results [4]. A good binarization algorithm is one that best preserves the foreground and omit as much background as possible.

Other topics to be considered in this study which forms part of the binarization algorithm is the calculation of a threshold value, pre-processing and post-processing techniques which aims to increase the performance of the binarization algorithm and finally the evaluation of the binarization method at the end of the study to determine the successfulness of the algorithm.

## 2.2 Existing Research

A considerable amount of binarization algorithms exist and they all have their advantages and disadvantages, some fit better for certain circumstances or for a certain type of image. No existing binarization algorithm is good to suit all types of images [6], but one can improve the result by combining algorithms [7], by adding pre-processing and post-processing techniques to the binarization algorithm such as noise removal [1], [6], or by careful experimentation and analysis on a very specific dataset in order to train the algorithm accordingly. One of the key objectives in binarization is thresholding, which is the calculation algorithm of the middle or "split" value used to discriminate the foreground from the background in the image. The algorithm would mark the pixels in a binary fashion and divide them into two categories, ON denoting the set of pixels representing the foreground or OFF denoting the set of pixels representing the background [10].

There are some research going on in historical document processing as seen in the Historical document Imaging and Processing Proceedings, the research contained in the Historical document Imaging and Processing Proceedings documentation will probably be of good use during this project, but they mostly focus on character recognition or word spotting of foreign characters such as Arabic and Chinese writing. Another similar application which is also more popular is the identification and characterization of the writer or author of historical documentation [9]. It seems that little focus is placed on the binarization of historical handwritten documents resulting in insufficient research [5] which provides the opportunity to do more research in this field.

Promising solutions are the combination of various binarization algorithms [11], as presented in the article with the title: Combination of Document Image Binarization Techniques, written by Su, Lu and Tan. and presented at the 2011 International Conference on Document Analysis and Recognition. It is probable that this approach would be the most successful since it was tested on historical documentation. Gatos's binarization algorithm also demonstrated promising results in an article named Adaptive Degraded Document Image Binarization, written by Gatos, Pratikakis and Perantonis[3], although these images weren't historical but degraded in a way that is comparable to historical documentation. The techniques presented in Binarizaing Complex Scanned Documents by Lins and Almeida [6] might also be considered since these techniques handle problems similar to those one encounter in historical document image processing. Other proven binarization algorithms includes Souvola's and Otsu's algorithms but have only been proven successful in good circumstances. The algorithms of Chen and Wang will also be considered during this research.

Existing Research also shows that pre-processing and post-processing techniques can greatly improve the binarization results [5], [7].

### 2.3 Rationale

The reason for this research project is that there is not many successful binarization algorithms capable of performing the task this problem requires, most algorithms are adequate in only the perfect circumstances such as on good quality documents [5], or adapted for a specific set of image types, but this project deals with degraded and historically written documentation which is very daunting to read even for human eyes. Other reasons accountable for the lack of success are the order of variations in degraded documentation, there are simply too many variations to generalize all under one class for one algorithm [7].

### 2.4 Problem Statement

The problem statement is to determine and implement the binarization algorithm which best preserves the quality of the document images, specifically historical handwritten text on degraded paper. To preserve the quality of a document image is to extract the useful information with high accuracy from the image while omitting the background and irrelevant information. See *Figure 1* for an example of the document images to be binarized.

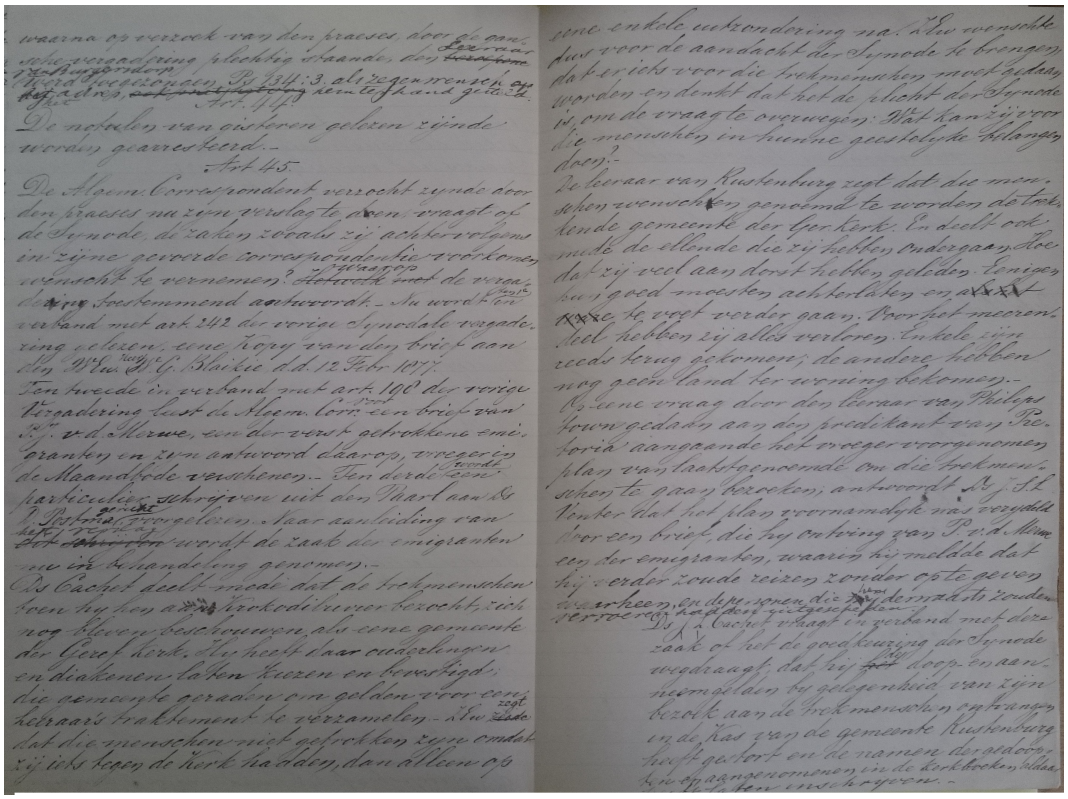

*Figure 1: Example of document image to be binarized*

## 3 Research Aims and Objectives

The aim of this study is to propose a binarization algorithm that bests differentiate the foreground(text, tables and figures) from the background(paper and stains) in historical handwritten documentation, specifically on the dataset which is the minutes of the early synod meetings of the Reformed Churches in South-Africa.

The objectives to accomplish in order to achieve the above aim is to:

- Perform a literature study on thresholding techniques and binarization algorithms for grey scale images, handwritten documents and historical documentation. A small literature study will also be done on pre-processing and post-processing techniques for document binarization, it will be small because it is not very common to implement these extra processing techniques. It is also necessary to perform a literature study on the evaluation of document image binarization.

- Familiarization with MathWorks Matlab and/or the Matlab Image Processing Toolbox which will be used to test and analyse the thresholding, pre-processing and post-processing techniques and binarization algorithms, the design and implementation of the binarization algorithm will also be done in the Matlab environment at the end of the study.
- Analysis of the performance of existing thresholding techniques and binarization algorithms on the historical handwritten documents dataset. Experimenting with pre-processing and post-processing techniques will also be done to achieve maximum performance from the binarization algorithm.
- Improve existing binarization algorithm(s), combine existing binarization algorithms, incorporate other image processing techniques into existing binarization algorithms or design a new binarization algorithm if existing solutions yield totally inadequate binarization results.
- Evaluate and implement the final binarization algorithm(s) of choice.

# 4 Procedures and Methods

This section will discuss the research methodology as well as the procedures and methods that will be followed to finally conclude this research. The research methodology will shape the process this project will undergo and methods chosen when experimenting with solutions.

## 4.1 Research Paradigm

Research is defined as "the creation of new knowledge using an appropriate process, to the satisfaction of the users of the research" [8]. There are numerous reasons for doing research, the research conducted during this project has three reasons behind it, the first is to add to the body of knowledge. Since there aren't many successful binarization algorithms for the binarization of historical handwritten document images, this research will aim to provide the scientific community with an algorithm which is tested and proven to be the most effective for the binarization of historical handwritten document images. The second reason is to solve a problem, as stated in the problem description section this project is motivated to digitalize historical documentation and to solve the problem of the degradation and loss of information. The third reason behind this research is to come up with a better way, a better way of binarizing historical handwritten document images. The existing methods or binarization algorithms does not suffice as explained

in the problem description section.

Research usually concludes into something useful to someone. This research will result in a part-product, part-theory and a critical analysis. The theory may be completely new or may be a reinterpretation of an existing theory, this research might use existing theories or techniques in image processing and apply it in a different context, the word "theory" is used because it has yet to be tested and accepted in practice as well. The critical analysis will be the analysis of the binarization algorithms and the various techniques they incorporate.

A research paradigm is a way of thinking about aspects of the world when doing research, three research paradigms exists, positivism, interpretivism and critical research of which positivism is the most commonly used paradigm. Interpretivism is a study where one tries to create and provide an understanding of how certain factors are related in a social context, therefore no hypothesis is required in the interpretivism philosophical paradigm. There exist no acceptable success criteria for the interpretivism paradigm. Critical research is concerned with the contradictions in the modern world and to aid in the elimination of alienation and domination. This research project will follow a positivism philosophical paradigm, which is an approach to research in the natural sciences. The positivism philosophical paradigm is also know as the scientific method [8] or the hypothetico-deductive method.

This research will follow the hypothetico-deductive methodology as explained by Dodig-Crnkovic, 2009. This methodology is found suitable for this research project because it is a scientific method for the study of natural sciences, and since this project mainly entails image processing which is a branch of computer science which originates from the fields of mathematics and logic which is both fields of the natural sciences [12]. "Computer science is the study of computational processes and information structures, including their hardware realizations, their linguistic models, and their applications" [12]. The scientific or hypothetico-deductive methodology is chosen as the research methodology. This methodology excels in this type of research because it evaluates the solutions through experimenting or the construction of prototypes, which is algorithms in the case of this research project [2].

The hypothetico-deductive methodology works on the assumptions that the world is ordered,

regular and not random which one can investigate in an objective manner. Everything can be investigated objectively because everything is independent of human cognition. The hypothetico-deductive methodology can never be proven true, a scientist can only predict what should happen (with a certain probability) by using the knowledge concluded from a series of experiments [8], but a good theory must always have the possibility of falsification. If the theory is not falsifiable, it should not be trusted, because a theory becomes trustworthy once it has been tested many times with failure to falsify.

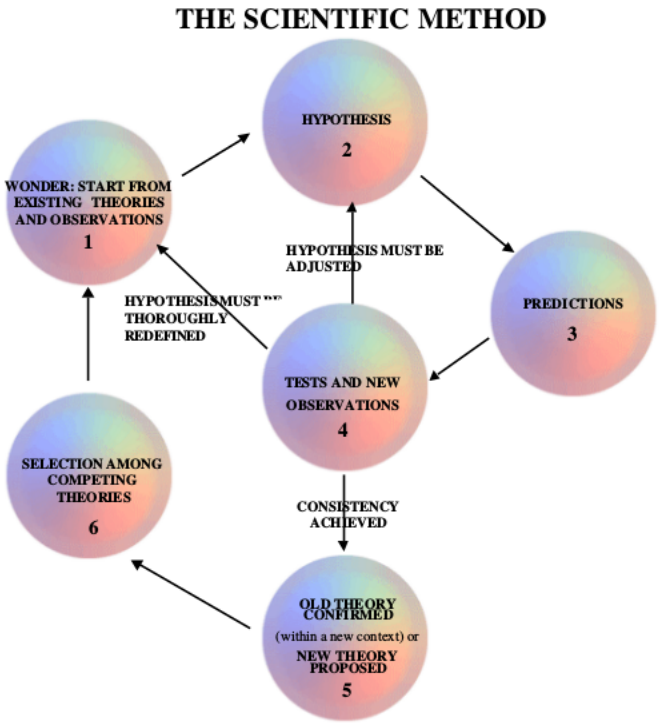


*Figure 2: The Scientific Method* [2]

The hypothetico-deductive methodology follows a recursive series of six steps as explained below and shown in *Figure 2*:

1. Provide a research question or problem statement which can be answered or solved in terms of existing knowledge or one that requires a new theory.

2. Formulate a hypothesis.

3. Deduce consequences and make predictions.

4. Conduct experiments to test the hypothesis objectively, existing science must accommodate the hypothesis. If the hypothesis contradicts existing science, the hypothesis must be refuted and careful testing must follow. Steps two through three is repeated until consistency is obtained.

5. If consistency is obtained through enough iterations, the hypothesis becomes a confirmed theory.

6. The new theory is compared to existing theories.

Research methods of this study will include the following subsections.

## 4.2 Algorithm Performance Analysis

Performance analysis of algorithms will be done by experimenting and prototyping of the algorithms and by comparing and evaluating the binarization results. The algorithms will be tested on the minutes of the early synod meetings of the Reformed Churches in South-Africa dataset by using Matlab.

## 4.3 Algorithm Design

Modification of existing algorithms or the design of a new algorithm, if opportune, will also be done on Matlab as well as the implementation of the final algorithm that will be used to binarize the images in the previous mentioned dataset.

## 4.4 Binarization Result Evaluation

The results of the binarization will be evaluated by using the research done on evaluation as well as by comparison and inspection by the study leaders and/or other computer scientists.

## 4.5 Artefact

The resulting artefact will be a binarization algorithm with a step by step explanation of the algorithm, the Matlab code for the algorithm will also be provided as well as examples of the results produced by the algorithm to determine the successfulness of the algorithm. A prototyping development method will be used because a series of different prototypes(Matlab code and results) will be compared against each other and modifications to the algorithms will happen iteratively

until the best performing algorithm is found.

This project will be very similar to the project of Jacques B. Barnard's, the only difference is that the data of Barnard's project consists of printed text while the data of this project consists of handwritten text, with both texts containing similar content.

# 5 Approach to Project Management and Project Plan

As stated in the previous chapter, this project will follow an iterative prototyping development methodology. The work needed to be done for this project is defined in a series of deliverables that has to be completed and submitted on predefined dates as stated in the study guide for this module, ITRI 671. Each deliverable has its own objectives and requirements to be met. The work flow for this project will be presented as a Gantt chart created with Planner 0.14.6, a project management application for the GNOME desktop. This Gantt chart will be used throughout the research term to maintain the project's time schedule.

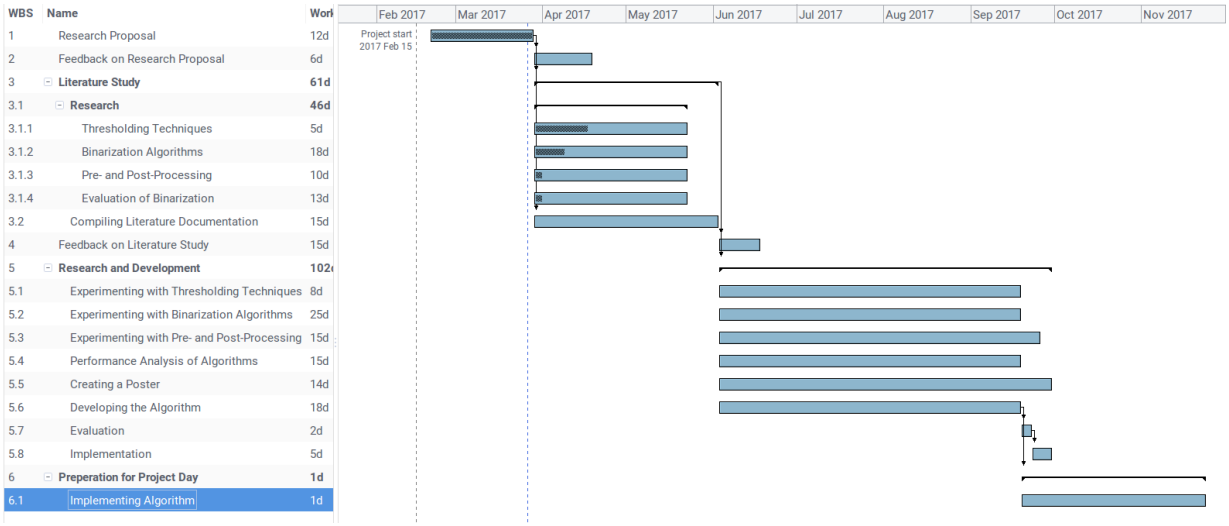


*Figure 3: Project Gantt Chart*

This project commenced on the 15$^{th}$ of February 2017 and will finish on the 23$^{rd}$ of November 2017. The following milestones are important for the success and time management of this project. Submission of the project proposal, this document, will be on the 27$^{th}$ of March 2017 at 17:00. Students get feedback from supervisors about their project proposals on the 18$^{th}$ of April 2017.

The literature study will be submitted on the 2nd of June 2017. All students will get feedback from supervisors about their literature study on the 17th of July 2017. The artefact or final product that this research study has concluded to will be demonstrated on the 29th of September 2017 along with a poster summarizing the research study. The submission of the complete set of documentation will be on the 23rd of October 2017. There will be a project day on the 23rd of November 2017 where all project will be on display as well as presentations about the projects.

Fortunately this research project requires no additional financing and all equipment and software that'll be necessary for this research project to succeed are paid for and available. There are no additional costs to research since the university already has loads of documentation and information, and the natural sciences library will purchase any additional information with the inter library loans if requested.

### 5.1 Scope

As stated in the Introduction chapter this project is only concerned with the binarization aspect of the broader aim of this project. The binarization will consist of the calculation of a threshold value and might include pre-processing and post-processing techniques as well. The binarization algorithm will be specifically chosen, improved or designed and implemented only for the minutes of the early synod meetings of the reformed churches in South-Africa, which is handwritten text on degraded paper but may not perform as well on other types of documentation.

### 5.2 Risks

The only risk this projects entails is the risk of damaging the historical archives of the theological library while scanning the documentation. Some precautions are taken to keep this risk to a minimum such as wearing cotton gloves when handling the documentation and using a tweezers to lift a page. The documentation is also handled with extra care when paging and scanning and folding the pages is prohibited.

## 6 Development Platform, Resources and Environments

The environment will be strictly scientific because this project is of no interest in an economical environment nor has it economic purposes. The final implementation will only be used once,

the results or processed data will be presented to users or community thereafter. The research, experience and knowledge gathered will be well documented and shared to become of use to other computer scientists who seeks to achieve similar goals for the sole purpose to broaden the knowledge of the scientific community.

A personal computer(MSI GE62VR-6RF Apache Pro Laptop) will be used for processing since the processing does not require an industrial level of processing. The computer has an Intel Core i7 6700HQ processor with 4 cores and 8 threads running at a clock speed of 2.6GHz utilizing 8GB of DDR4 RAM and a nVidia GeForce GTX 1060 with 6GB of GDDR5 RAM. The screen on which the binarized images will be compared on is a 15" MSI LCD IPS panel with a resolution of 1920x1080. The computer and hardware are of sufficient quality and has the performance capabilities suitable to satisfy all requirements necessary in order to accomplish the goals of this project. Processing will be done via MathWorks Matlab and/or the Matlab Image Processing Toolbox since the software will only be used for scientific purposes and not for consumer nor commercial use. Matlab started out as a matrix laboratory many years ago, today it is a mathematical tool for mathematicians and scientists used to perform complex calculations at a large scale. Matlab contains all the functions and features necessary to overcome all objectives required to achieve the aim of this project. The personal computer is running an operating system called Zorin OS Ultimate, with generic Linux kernel 4.8.0-41 and Gnome/Zorin Desktop Environment. LaTeX in conjunction with TeXmaker will be used for all documentation because of its simplicity, accuracy, neatness and the ease of use of TeXmaker, the resulting document will be presented in PDF format because PDF is used very widely and highly compatible with many systems.

Other resources that'll be used includes the MathWorks Matlab help documentation and numerous books on Matlab such as Introduction to Matlab and Simulink - A project Approach 3rd Ed., 2008 by Beucher and Weeks; Matlab - A Practical Introduction to Programming and Problem Solving 3rd Ed., 2013; Matlab for Dummies, 2015 by Sizemore and Mueller; Matlab Programming Fundamentals R2015a, 2015; and Practical Image and Video Processing Using Matlab, 2011 by Oge Marques.

# 7 Ethical and Legal Implications

This project has only one ethical implication because no people not related to this project will be affected during the research term. The only ethical implication is that the research conducted is original and that no plagiarism is involved. The only legal implications are the use of software in a legal fashion. The MathWorks Matlab software and all toolboxes are used with an academic license obtained from the North-West University. Word processing software being used for the documentation such as this document are the LaTeX libraries and the TeXmaker LaTeX interface which is being used under the free GNU General Public License version 2. The operating system being used is Zorin OS Ultimate which is paid for in full by the author of this document. Any other operating systems which might be used are other Linux distributions such as Manjaro which is free to use and also released under the GNU General Public License version 2.

# 8 Provisional Chapter Division

## 8.1 Introduction

The introduction will include a description of the project, the aim of the project along with the objectives to be completed in order to achieve the aim of this project. The introduction chapter will also include a background a a rationale. The methods of investigation will also be discussed in this chapter.

## 8.2 Research Methodology

The research methodology chosen for this project will be discussed as well as the reasons for choosing this methodology.

## 8.3 Literature Study

This chapter will contain the literature study on the research topics at hand, namely a small part on the research methodology, thresholding techniques, followed by binarization algorithms as well as pre-processing and post-processing techniques and finally the evaluation of binarization algorithms in order to determine the successfulness of the proposed algorithms.

## 8.4 Threshold Calculation

The method of calculation for the threshold will be discussed here as well as the justifications for using the proposed calculation(s).

## 8.5 Evaluation of Binarization Algorithms

Binarization algorithms will be evaluated in this chapter and compared against each other. The combination of binarization algorithms will be evaluated in this chapter as well.

## 8.6 Pre- and Post-Processing Techniques

Pre- and post-processing techniques that forms part of the binarization algorithm will be explained in this section.

## 8.7 Design/Implementation of Binarization Algorithm

The implementation of the proposed binarization algorithm or combination of algorithms will be explained in this chapter. The design of a new binarization algorithm will also be explained if necessary and possible.

## 8.8 Results

The result of the proposed algorithm(s) or solution will be demonstrated in this chapter in order to determine the successfulness of the algorithm. The results will either be presented as binarized images for visual inspection or in a table containing statistical data to compare the successfulness of the algorithm to other existing algorithms.

## 8.9 Conclusion

The conclusion derived from the research done will be presented in this chapter as well as future research and improvements needed to be done.

## 8.10 Reflection

A summary of the research study accomplishments will be given in this chapter along with the student's views and reflection about this project.

## References


[1] Mudit Agrawal and David Doermann. Stroke-like pattern noise removal in binary document images. In *2011 International Conference on Document Analysis and Recognition (ICDAR)*, pages 17–21. IEEE, 2011.

[2] Gordana Dodig-Crnkovic. Scientific methods in computer science. In *Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde, Suecia*, pages 126–130, 2002.

[3] Basilios Gatos, Ioannis Pratikakis, and Stavros J Perantonis. Adaptive degraded document image binarization. *Pattern recognition*, 39(3):317–327, 2006.

[4] Nicholas R Howe. A laplacian energy for document binarization. In *2011 International Conference on Document Analysis and Recognition (ICDAR)*, pages 6–10. IEEE, 2011.

[5] Bency Jacob and SB Waykar. A survey on binarization of historical degraded documents. 2014.

[6] Rafael Dueire Lins, P Gabriel de França, and Marcos Martins de Almeida. Binarizing complex scanned documents. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 56–60. IEEE, 2015.

[7] Nikolaos Ntogas and Dimitrios Veintzas. A binarization algorithm for historical manuscripts. In *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*, number 12. World Scientific and Engineering Academy and Society, 2008.

[8] Briony J Oates. *Researching information systems and computing*. Sage, 2005.

[9] Jean-Marc Ogier and Karl Tombre. *Madonne: document image analysis techniques for cultural heritage documents*. na, 2006.

[10] Lawrence O'Gorman and Rangachar Kasturi. *Document image analysis*, volume 39. IEEE Computer Society Press Los Alamitos, 1995.

[11] Bolan Su, Shijian Lu, and Chew Lim Tan. Combination of document image binarization techniques. In *2011 International Conference on Document Analysis and Recognition (ICDAR)*, pages 22–26. IEEE, 2011.

[12] Allen B Tucker. *Computer science handbook*. CRC press, 2004.

# Appendix A: Research Proposal

# SUBJECT GROUP COMPUTER SCIENCE AND INFORMATION SYSTEMS

## Research Proposal for Honours project

The student and the supervisor must consult the *Manual for Postgraduate Studies* prior to writing the research proposal. The *Manual for Postgraduate Studies* explains in detail what is expected at each of the subheadings below. The proposal should not be longer than 5 pages.

The Subject Group requires that the research proposal will be submitted through the use of this form and in the format below. Please complete using a computer.

### 1 Student initials, surname and student number

| Initials | Surname | Student number |
|---|---|---|
| Z | Labuschagne | 23585137 |

### 2 Degree for which student is registered

B.Sc.(Hons) Computer Science and Information Systems

### 3 Name of supervisor

| Initials and surname |
|---|
| Prof. G.R. Drevin & Mev. S. Campher |

### 4 Proposed title

| | |
|---|---|
| Title | Binarization of the Minutes of the Early Synod Meetings of the Reformed Churches in South-Africa |

### 5 Problem statement and substantiation

Provide the theme and link with gaps in the literature and recent research in the area. Indicate the research question, its actuality and how the research will endeavour to answer the question.

Binarization in image processing is the process of discriminating the foreground from the background in an image by changing the foreground pixels or text to black and the background or paper pixels to white [5]. Binarization is also considered by many computer scientists as a critical step in the processing of images because it has the greatest impact on the quality grade of all the processing that follows such as document analysis [3], thus a good binarization result would greatly improve the results of page-segmentation, optical character recognition and any other subsequent processing [1], while a bad binarization result will cause any following processing techniques to yield inadequate results [2]. A good binarization algorithm is one that best preserves the foreground and omit as much background as possible.

A considerable amount of binarization algorithms exist and they all have their advantages and disadvantages, some fit better for certain circumstances or for a certain type of image. No existing binarization algorithm is good to suit all types of images [4], but one can improve the result by combining algorithms [5], by adding pre-processing and post-processing techniques to the binarization algorithm such as noise removal [1], [4], or by careful experimentation and analysis on a very specific dataset in order to train the algorithm accordingly.

There are some research going on in historical document processing as seen in the Historical document Imaging and Processing Proceedings, the research contained in the Historical document Imaging and Processing Proceedings documentation will probably be of good use during this project, but they mostly focus on character

recognition or word spotting of foreign characters such as Arabic and Chinese writing. Another similar application which is also more popular is the identification and characterization of the writer or author of historical documentation [6]. It seems that little focus is placed on the binarization of historical handwritten documents resulting in insufficient research [3] which provides the opportunity to do more research in this field.

The problem statement is to determine and implement the binarization algorithm which best preserves the quality of the document images, specifically historical handwritten text on degraded paper. To preserve the quality of a document image is to extract the useful information with high accuracy from the image while omitting the background and irrelevant information.

## 6 Research aims and objectives

Provide the different general as well as the specific aspects which will form part of the research.

The aim of this study is to propose a binarization algorithm that bests differentiate the foreground(text, tables and figures) from the background(paper and stains) in historical handwritten documentation, specifically on the dataset which is the minutes of the early synod meetings of the Reformed Churches in South-Africa.

The objectives to accomplish in order to achieve the above aim is to:

- Perform a literature study on thresholding techniques and binarization algorithms for grey scale images, handwritten documents and historical documentation. A small literature study will also be done on pre-processing and post-processing techniques for document binarization, it will be small because it is not very common to implement these extra processing techniques. It is also necessary to perform a literature study on the evaluation of document image binarization.
- Familiarization with MathWorks Matlab and/or the Matlab Image Processing Toolbox which will be used to test and analyse the thresholding, pre-processing and post-processing techniques and binarization algorithms, the design and implementation of the binarization algorithm will also be done in the Matlab environment at the end of the study.
- Analysis of the performance of existing thresholding techniques and binarization algorithms on the historical handwritten documents dataset. Experimenting with pre-processing and post-processing techniques will also be done to achieve maximum performance from the binarization algorithm.
- Improve existing binarization algorithm(s), combine existing binarization algorithms, incorporate other image processing techniques into existing binarization algorithms or design a new binarization algorithm if existing solutions yield totally inadequate binarization results.
- Evaluate and implement the final binarization algorithm(s) of choice.

## 7 Basic hypothesis (where applicable)

N/A

# 8 Method of investigation

## 8.1 Literature study

Provide an indication only of which literature will be used in the study with a few key references. A summary of the literature is not required here.

Literature to be considered during research:

**Research Methodology:**

- Scientific Methods in Computer Science (Dodig-Crnkovic).
- The Generality of Hypothetico-Deductive Reasoning: Making Scientific Thinking Explicit (Lawson, 2000).
- Hypothetico-Deductive Confirmation (Sprenger, 2011).

**Thresholding:**

- Document Image Analysis (O'Gorman & Kasturi, 1997).

**Document Image Binarization:**

- Madonne: Document Image Analysis Techniques for Cultural Heritage Documents (Ogier & Tombre, 2006).
- A Binarization Algorithm for Historical Manuscripts (Ntogas & Ventzas, 2008).
- A Survey on Binarization of Historical Degraded Documents (Jacob & Waykar, 2014).
- Efficient Binarization of Historical and Degraded Document Images (Gatos, Pratikakis & Perantonis, 2008).
- Binarization of Historical Document Images Using the Local Maximum and Minimum (Su, Lu & Tan, 2010).
- Using Binarization Method Rebuilding of Historic Document Images (Sireesha, Haritha & Manasa, 2016).
- Why Multiple Document Image Binarizations Improve OCR (Lund, Kennard & Ringger, 2013).
- A Tool for Tuning Binarization Techniques(Sokratis & Kavalieratou, 2011).
- A Laplacian Energy for Document Binarization(Howe, 2011).
- Stroke-Like Pattern Noise Removal in Binary Document Images(Agrawal & Doermann, 2011).
- Binarizing Complex Scanned Documents (Lins, Silva & de Almeida, 2015).
- MRF Based Text Binarization in Complex Images using Stroke Feature (Wang, Shi, Xiao & Wang, 2015).
- Combination of Document Image Binarization Techniques (Su, Lu & Tan, 2011).

**Evaluating Binarization Techniques:**

- Information Retrieval (van Rijsbergen, 1979).
- Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation (Powers, 2007).
- An Objective Evaluation Methodology for Document Image Binarization Techniques (Ntirogiannis, Gatos & Pratikakis, 2008).
- Performance Measures for Information Extraction (Makhoul, Kubala, Schwartz & Weischedel, 1999).

- An introduction to ROC analysis (Fawcett, 2005).
- When is a Problem Solved? (Lopresti & Nagy, 2011).
- Quality evaluation of ancient digitized documents for binarization prediction (Vincent, Nicholas, Vialard & Domenger, 2013).

### 8.2 Methods of investigation

The proposed design, data acquisition, procedures, data processing, funding sources (but not a budget), mathematical methods, computer methods, etc.

**Research Methodology**

This research will follow an hypothetico-deductive methodology as explained by Dodig-Crnkovic, 2009. This methodology is found suitable for this research project because it is a scientific method for the computer science field. This methodology excels in this type of research because it evaluates the solutions through the construction of prototypes, which is algorithms in the case of this research project [3].

**Algorithm Performance Analysis**

Performance analysis of algorithms will be done by experimenting and prototyping of the algorithms and by comparing and evaluating the binarization results. The algorithms will be tested on the minutes of the early synod meetings of the Reformed Churches in South-Africa dataset by using MATLAB.

**Algorithm Design**

Modification of existing algorithms or the design of a new algorithm, if opportune, will also be done on MATLAB as well as the implementation of the final algorithm that will be used to binarize the data in the previous mentioned dataset.

**Binarization Result Evaluation**

The results of the binarization will be evaluated by using the research done on evaluation as well as by comparison and inspection by the study leaders and/or other computer scientists.

## 9 Provisional chapter division

Here it should be clear that there was proper reflection on the appearance of the final product (mini dissertation). Provide provisional titles of the various chapters, with a brief outline of the planned content of each.

1. **Introduction**

The introduction will include a description of the project, the aim of the project along with the objectives to be completed in order to achieve the aim of this project. The introduction chapter will also include a background a a rationale. The methods of investigation will also be discussed in this chapter.

2. **Research Methodology**

The research methodology chosen for this project will be discussed as well as the reasons for choosing this methodology.

3. **Literature Study**

This chapter will contain the literature study on the research topics at hand, namely a small part on the research methodology, thresholding techniques, followed by binarization algorithms as well as the evaluation of binarization algorithms in order to determine the successfulness of the proposed algorithms.

4. **Threshold Calculation**

The method of calculation for the threshold will be discussed here as well as the justifications for using the proposed calculation(s).

5. **Evaluation of Binarization Algorithms**

Binarization algorithms will be evaluated in this chapter and compared against each other. The combination of binarization algorithms will be evaluated in this chapter as well.

6. **Design/Implementation of the Binarization Algorithm**

The implementation of the proposed binarization algorithm or combination of algorithms will be explained in this chapter. The design of a new binarization algorithm will also be explained if necessary and possible.

7. **Results**

The result of the proposed algorithm(s) or solution will be demonstrated in this chapter in order to determine the successfulness of the algorithm.

8. **Conclusion**

The conclusion derived from the research done will be presented in this chapter.

9. **Reflection**

A summary of the research study accomplishments will be given.

## 10 Literature references

Provide complete references to the literature referenced to in this proposal only.


[1] Mudit Agrawal and David Doermann. Stroke-like pattern noise removal in binary document images. In 2011 International Conference on Document Analysis and Recognition (ICDAR), pages 17–21. IEEE, 2011.

[2] Nicholas R Howe. A laplacian energy for document binarization. In 2011 International Conference on Document Analysis and Recognition (ICDAR), pages 6–10. IEEE, 2011.

[3] Bency Jacob and SB Waykar. A survey on binarization of historical degraded documents. 2014.

[4] Rafael Dueire Lins, P Gabriel de França, and Marcos Martins de Almeida. Binarizing complex scanned documents. In 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pages 56–60. IEEE, 2015.

[5] Nikolaos Ntogas and Dimitrios Veintzas. A binarization algorithm for historical manuscripts. In WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering, number 12. World Scientific and Engineering Academy and Society, 2008.

[6] Jean-Marc Ogier and Karl Tombre. Madonne: document image analysis techniques for cultural heritage documents. na, 2006.


..... ..... .....

Student Supervisor Date