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	Z	Surname	Labuschagne	Student number	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