Machine Learning

Project Proposal

For

Binarization of Historical Documents

Prepared by   
  
Faizan Zafar  
Saad Raza  
Syed Wahab Aftab  
Umair Khalid

BESE-5B | SEECS

27th March, 2017

Table of Contents

Table of Contents ii

Revision History iii

1.0 Introduction 4

2.0 Problem Statement 4

3.0 Solution 4

**3.1 Scope of the project………………………………………….……………………….4  
3.2 Novelty………...………..…………………………………………………………….5  
3.3 Team Members’ Roles…………………………..……………………………….......5  
3.4 Timeline of Completion………..…………………………………………………….6**

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Authors** | **Date** | **Reason For Changes** | **Version** |
| **Umair Khalid**  **Syed Wahab Aftab**  **Saad Raza**  **Faizan Zafar** | 27/03/2017 | Project Proposal: Final Draft | 1.0 |

# Introduction

Historical documents and ancient manuscripts are an integral part of a society’s culture, science, civilization and identity. The need to preserve and understand them has always been of high importance. Degradations and impurities have led to these documents being almost unreadable. The media used is parchment or papyrus which is subject to deterioration. In present, computers can process a large amount of images of these documents in multimedia formats for future analysis and storage. Although current systems can store all these images, there is no specific system that is capable to retrieve relevant information efficiently and to extract knowledge from them.

Binarization of ancient document images is a process of removing unrelated artefacts and background noise. Image Binarization is essential not only to document image analysis, but also to significantly improve the overall performance as poor Binarization will result in poor information segmentation of the original characters and additional noise could be added to the image. Therefore, determining proper Binarization techniques is a key significant factor in achieving promising results from document image analysis.

The purpose of this document is to define the problem with these historical documents and provide a feasible solution based on our expertise. The solution has some detailed aspects which we have discussed so the reader of this document can fully encompass the operations and limitations of the project.

# Problem Statement

Scanned documents are being used in all types of fields. The fact that they are not always clear and sometimes give poor quality images is a let-down, this makes it difficult to read and understand them. Also they don’t provide search functionality. As most of the books are now scanned copies, it makes it harder to find a topic of interest manually without machine assistance.

# Solution

Our solution will focus on Binarization, we would convert our image to greyscale and then instead of doing global Binarization we would make an algorithm which would find the appropriate threshold and binarize. For information segmentation we will make our algorithm in such a way that it would be able to distinguish and detect different characters and then recognizing them automatically. The project will cater all who use scanned documents in any shape or form i.e in study, research or office work etc. The project will be made on Pycharm in python language with OpenCv library.

# Scope of the project

The scanning of important documents is a very common task nowadays but accurate and clear scanning is an issue which not catered in common tools we use to scan them. Our project will enable a common user to scan the document without worrying about the clarity and accuracy of the scanning because of our algorithm which will take care of the distortions or any other problems in the document because of its age or condition. The application will be applicable to any daily use home computers.

# Novelty

# Binarization and Information Segmentation are both not unique concepts as they have been done many times before, what we want is to make a system which takes an image binarize it and then further segment it. What it does is that it provides dual functionality for documents to the user, we know that almost all documents would find Binarization and segmentation useful. So, we grabbed two concepts and combined them into one and presented a different system.

# 3.3 Team Members’ Roles

* **Faizan Zafar**

Responsible for selecting the appropriate platform for the implementation phase and then working on it within the scope of the end semester project.

* **Syed Wahab Aftab**

Responsible for the selecting the appropriate Binarization technique in the scope of the end semester project after extensive research.

* **Umair Khalid**

Responsible for the documentation of the whole end semester project providing benchmarks, progress reports and goals & objectives for each phase of the process.

* **Saad Raza**

Responsible for retrieving the appropriate test data for the project and applying the concepts of machine learning for text recognition and information segmentation.

These are specific roles and responsibilities of the team members. All team members will be equally contributing in the making of a final product, usually during the implementation phase.

# 3.4 Timeline of Completion

Following is a division and tentative timeline of the project’s development lifecycle:

* **Requirements Elicitation and Analysis**

27th March 2017 to 2nd April 2017 *(One week)*

* **Approach selection and Analysis**

3rd April 2017 to 10th April 2017 *(One week)*

* **Implementation**

10th April 2017 to 15th May 2017 *(Five weeks)*

* **Testing against data set**

15th May 2017 to 22nd May 2017 *(One week)*