READABILITY ENHANCEMENT OF HISTORICAL DOCUMENT IMAGES

MSC IN COMPUTER SCIENCE

D.D.G.A.C.W. SAPARAMADU

Department of Computer Science and Engineering

University of Moratuwa

12/2007

READABILITY ENHANCEMENT OF HISTORICAL DOCUMENT IMAGES

D.D.G.A.C.W. SAPARAMADU

This Dissertation was submitted to the Department of Computer Science and Engineering of the University of Moratuwa in Partial Fulfillment of the requirements for the Degree of MSc in Computer Science

Department of Computer Science and Engineering
University of Moratuwa
12/2007

Declaration

I, D.D.G.A.C.W. Saparamadu hereby declare that the work included in this dissertation in part or whole has not been submitted for any other academic qualification at any institution.

Dr. Chatura De Silva C.W. Saparamadu

Supervisor

Abstract

The Department of National Archives of Sri Lanka posses a huge collection of historical documents and those can be considered as a national heritage as it provides documental evidences for country's history. The documents anyway are subjected to many degradations due to natural aging, acidification, manual handling and lack of maintenance. The department is in the process of digitizing the documents as a preservation strategy and faced the problem of illegible document images after the digitization process. This research is conducted to find the solutions for this problem and particularly paid attention on ink-bleed-through problem, which occurs in aged, double sided documents.

We have evaluated many traditional image enhancement techniques like contrast stretching, histogram equalization and also paid attention on some other novice technologies like *k*-Means clustering and Principle Component Analysis. In order to conduct the experiments, we have used Image Processing Toolkit of Matlab and written an application to explore the performance of algorithms in depth. Further we have extended our experiments even to many different color spaces and results were analyzed in detail. This flexible approach has helped us to practically evaluate an existing algorithm with actual document images and to understand the pros and cons of each approach.

We found that some traditional image enhancement techniques are contributing positively on historical document image enhancement, while some others are not. The application of commonly used image registration based techniques are found not feasible in real-time environment and certain other techniques like color segmentation seems capable of providing a reasonable output within a very short time. *k*-Means clustering coupled with Principle Component Analysis is yet another combined solution and its performance and the limitations are described.

Based on the analysis, we have reached certain conclusions and some recommendations were made for a real-time historical document image enhancement strategy.

Acknowledgements

Throughout this research, the Department of National Archives of Sri Lanka has extended their fullest corporation and without it, this work not had been successful. The Director of National Archives, Dr. Saroja Wettasinghe and Planning Assistant, Mr. Chaminda Mihindukula Suriya were always accessible and their cordial support and guidance helped immensely to better understand the problem. They were able to supply us with necessary document images within a very short time and put lot effort in finding many ink-bleed-through related examples as that was the most challenging task that we had to address.

Project supervisor, Dr. Chatura De Silva has shown lot of flexibility in consulting him either personally, through emails or even via phone as it saved lot of time and effort and led us in the correct direction. Once we were in desperate need of certain research material, his intervention helped to contact authors of the material and get certain things clarified. I would really like to appreciate his support extended to date from the day the project has initiated.

Dr. Damith C. Rajapakse of National University of Singapore is appreciated warmly as he personally intervened in getting certain research material and Dr. Sanath Jayasena also helped immensely in this regard. Another big thank is deserved by Dr. Sanath Jayasena for his relentless efforts on scheduling progress meetings and guiding us on a methodical project progress.

Table of Contents

List of Figures

List of Tables

List of Symbols, Notations, Abbreviations and Acronyms

PCA - Principle Component Analysis

HSI - Hue-Saturation-Intensity color space

RGB – Red-Green-Blue color space

CIE - International Commission on Illumination

Lab - CIE 1976 (L* a* b*) color space or CIELAB

CMYK – Cyan-Magenta-Yellow-Black color space