Generation of Images of Historical Documents by Composition

Carlos A.B. Mello

Faculdade Santa Maria, Recife (PE), Brazil

cabm@netpe.com.br

Rafael D. Lins

Departamento de Eletrônica e Sistemas, UFPE, Recife (PE), Brazil

rdl@ee.ufpe.br

Abstract

This paper describes a system for efficient storage, indexing and network transmission of of historical documents. First, documents are decomposed into their features such as paper texture, colours, typewritten parts, handwritten parts, pictures, etc. Common components are factores out. Document retrieval forces the re-assembling of the document, synthetising an image visually close to the original document. The information needed to build the final image occupies, in average, 2 Kbytes performing a very efficient compression scheme.

1 Introduction

The work reported herein is part of the Nabuco Project[5][10] for preservation and broadcasting of the letters and documents from Joaquim Nabuco[1]'s bequest. The file is composed of almost 6,500 documents from the end of the nineteenth century, totalizing over 30,000 pages.

The Nabuco Project is developed by the Federal University of Pernambuco jointly with the Joaquim Nabuco Foundation (a social science research centre), both in Recife (Brazil). Documents are digitized in true colour with 200 dpi resolution and stored in JPEG [9] file format with 1% loss. Even in this format each image of a document reaches, in average, 380 Kb.

A processing environment was envisaged to extract the basic features of documents, which allows for later image re-assembling. The extraction of documents is performed by the blocks presented in Figure 1.

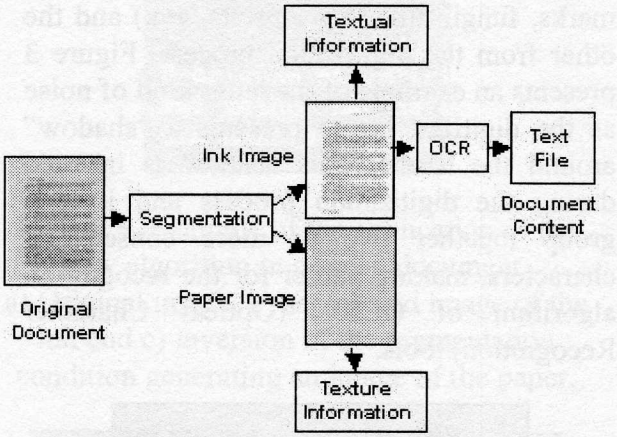

Figure 1. Block diagram for the data extraction of the paper and text image.

In order to obtain satisfactory results, several new algorithms were developed within the scope of the project and are described herein.

Ink and paper segmentation is not always a simple task in this kind of image. In some documents, the ink has faded; some of the others were written on both sides of the paper and the ink transposed the document presenting back-to-front interference. A conversion into a monochromatic version of this kind of documents using a nearest colour threshold algorithm [3] does not achieve high quality results as can be seen in Figure 2

[1] Brazilian statesman, writer, and diplomat, one of the key figures in the campaign for freeing black slaves in Brazil, Brazilian ambassador to London (b.1861-d.1910)