Habilitation thesis

Analysis, Processing, Information Retrieval and Storage of Document Images

Prof. Dr. Eng. Costin-Anton Boiangiu

SUMMARY

During the post-doctoral period, attention was turned towards the field of document image analysis.

Section 2.1, dedicated to the undertaken research, follows the normal flow of processing the documents, from initial scanning to conversion into digital form, while keeping the original look and feel.

The section is split into 6 chapters dedicated to the central theme and one dedicated to unrelated research.

Description of the research begins with the acquisition of the document images from analogic documents, having in mind the preservation of the colors and clarity of the original document. For achieving this task we propose methods for precision measuring and calibration.

Once digitized, the images undergo a phase of blur correction by deconvolution and another of conversion to grayscale, with minimal loss of useful information, the latter being necessary in the following stages of binarization. The same algorithms can be applied for printing a document on a mono-color printer.

For sending a simpler image to the analysis phase, various thresholding algorithms (global, local, and adaptive) follow, culminating with an important result of processing that does not require user intervention.

The last correction stage is the detection and correction of the rotation caused by the scanning process. Because it can happen that the size of the scanned document exceeds the scanning device, merging methods for the document fragments must also be considered. Fragments may or may not contain common borders. Still, we correct artifacts at the extremities of the page and crop the images, following an automatic detection of the printing area.

The correction and conversion phases are followed by the actual interpretation. Here are discussed detection and classification algorithms. The chapter starts with white space or any form separators detection. It continues with a modified Hough transform for detecting digital lines from tables in documents. In terms of content we present detection of text lines, even hand-written and strongly curved ones. Algorithms and methods include detection and

merging of fragmented letter elements. The next process analyses texture and measures font characteristics present in the document, boldness, inclination and size. Using the methods presented in the last section we can segment the images in areas that contain only text and order them to be sent to OCR. Several methods of segmentation, labeling and ordering the regions are discussed. The chapter ends with an important result in Computational Geometry - the Beta Shape, used for optimally encapsulating the text in non-overlapping polygonal regions in order not to process the same text twice.

Chapter 6 is dedicated to the actual reading of the content through OCR. We propose a step of post processing using dictionaries. Here are also presented formats for saving document image types of data, after being passed through all the processing stages. One of the important formats proposed is the MRC format which compresses each region differently to optimize the quality and size.

The last chapter contains research from other fields, which we may undertake in the near future. These are voting-based processing, introduced through voting-based image segmentation, and analysis of natural images, introduced by coin classification.

In terms of future work, we mentioned research directions in which we plan to continue (locality / globality and extending outside of image processing - electrical and financial markets -; additions to the proposed algorithms).

Advanced research will be carried during national or smaller projects. Here are exemplified TRISEMA (financial market prediction using the algorithms of image processing and neural networks) FINCRIS (measuring the financial crisis on the scale of the earthquake Richter scale) and ACID (altering the image on the monitor so that a person with eye deficiency can reconstruct it, without the need for glasses).

Activities with students will evolve towards competitive learning, the student team codes "battling" during the course, in a virtual environment, displayed on the projector. We also propose an online environment to conduct other such competitions (Geek Arena).

From the point of view of the subjects, they are being improved every year, considering student feedback including the students in research activities.