Automating Degraded Image Enhancement Processing (DIEP)

George Boutros
CiyaSoft Corporation
boutros@ciyasoft.com

1. Abstract

Documents recovered in the field by U.S. Government intelligence agencies have typically exhibited a myriad of image degradations related to environmental conditions, the use of primitive media, and careless handling. The sheer volume of these recovered documents overwhelms agencies faced with the task of translating Middle Eastern documents while impeded by an assortment of document image degradations.

Prior to performing Optical Character Recognition (OCR) and high-speed Machine Translation (MT) it is necessary that the degraded documents be evaluated and enhanced to remove many degradations. A high-speed enhancement paradigm is necessary to keep pace with the MT processing. Rapid processing for document image enhancement is inherently difficult due to the large number of algorithms that are required to identify, and remove, potential degradations.

The Degraded Image Enhancement Project (DIEP), conducted by CiyaSoft Corporation, will attempt to show that automated processing of sequences of intelligent high-speed algorithms can be developed and implemented into an FPGA platform capable of identifying and removing multiple degradations.

2. Goals

The DIEP Project focuses in general on developing intelligent software algorithms for processing Middle Eastern language degraded documents. In particular, methods of enhancing handwritten documents have been developed to prepare those documents for the OCR process.

The DIEP development goals are first to automate the selection of the specific degradation algorithms into sets of processes, related in a hierarchy of commonly encountered degradations, which are invoked simply by indicating a type of document. A migration of the software algorithms into VHDL code for implementation into a hardware platform would increase processing speed for each document by at least three orders of magnitude.

The final goal of the DIEP project is a produce a high-speed hardware platform that can accept batches of high resolution scanned degraded document images, and to produce as output the enhanced document images. The system should ideally be of sufficient speed to interface with MT processing.

3. Software Architecture

The DIEP meta-architecture is organized into two major functional partitions; an image layout understanding and analysis set of algorithms, and the degradation enhancement algorithms. The user GUI for the system establishes batch processing guidelines for groups of degraded documents.

Each degraded document is analyzed and the parametric information written to an associated configuration File. Typical file attributes would be extracting using the layout understanding and analysis algorithms to indicate if the document contains handwritten or printed text, languages that appear in the text, different fonts and styles, entity types, partition of entities, etc).

The Document attribute Format Spec (DAFS) is used for document decomposition. DAFS entities are defined to record content in a standardized set of entities. DAFS was originally developed under the Document Image Understanding (DIMUND) project funded by ARPA, and was employed for training and testing document image understanding tools. It is useful as a template for storing image content using one or more rectangular pieces of the image.

Hierarchical "parent", "child" and "sibling" relationships between entities permit levels of document decomposition, and DAFS permits assignment of a confidence level to each entities contents or properties (0 to 255). DAFS-B (binary) is used to store both Image and text binary format in one file.

The image segmentation and labeling algorithm separates text and graphics, identify font sizes in text blocks, and identifies printed text versus handwritten text. The language identification uses separate algorithms to identify handwritten text and printed text for Arabic, Farsi, and English. The algorithms for language identification operate differently to differentiate language when it is printed as opposed to handwritten. Printed text in Arabic and Farsi is challenging to discriminate without actually doing OCR. Handwritten styles in Farsi are distinctly different however from Arabic, and are easier to identify as Farsi.

Certain types of document combinations that are identified will “switch-on” selection of a sequence of "do-no-harm" enhancements for likely degradations that are commonly found in combination. For example, document attributes such as photocopied, FAX, handwritten, specific language (diacritics, character features), or historical/old can be used to "pre-select" a sequence of enhancements for likely degradations.

The second functional partition of DIEP algorithms, the enhancement algorithms are selected to operate on each parent entity within the document. The specific selection of enhancements is made based on the attributes of each entity.

DIEP enhancement algorithms are mostly adaptive, and intelligently remove the following degradations:

- Photocopy/FAX artifacts (backflash, poor contrast, noise, dirt, toner spread)
- Underlines (language specific)
- Paper folds and creases
- Stains/dirt

- Skewed documents (slanted text and graphics)
- Handwriting line slant – Arabic, Farsi, English (a highly localized effect)
- Poor image contrast (may be due to fading, poor media, or ink colors)
- Textured image backgrounds
- Graphics objects in the image
- Text character breaks (due to poor resolution, or poor writing instrument)
- Noise (transmission, speckle, random bits)
- Lined paper
- Irregular text character thickness
- Hole punch artifacts

4. Hardware Platform

Batch mode processing of many degraded documents is problematic when applied to the running of comprehensive sequences because it is much too slow. The most effective enhancement algorithms are computationally expensive, and inevitably too slow. Software solutions are therefore not practical for "bulk" processing when considering thousands or millions of archived documents.

Hardware platform solutions utilizing Field Programmable Gate Arrays (FPGAs) can dramatically increase processing time by as much as three orders of magnitude. Research comparing the same algorithm optimized on a RISC processor has been compared with FPGA, and shows speed improvements of 2-3 orders of magnitude (100x to 3000x times faster).

For example, a high resolution photocopied image that contains multiple degradations that include backflash, graphics objects and skewed text may take as long as 32 seconds to remove all degradations. An FPGA platform that is properly design to maximize speed of the algorithm processing can reduce that time to less than 10 ms.

FPGAs are flexible, reprogrammable devises that frequency boast “on-board” prcoessors, sucha s dual PowerPCs. They can be re-configured at run time from on-line memory, and multiple FPGAs can be ganged if necessary. The organization of FPGAs slices and macrocells can be designed not only to maximize a parallel pipelining of each algorithm, but also can process multiple algorithms in parallel.

The main goal of a software implementation is usually to minimize the number of iterations. In a hardware platform, our goals are to simplify the underlying operations as much as possible in order to speed up the calculations and to be able to provide more parallelism. A simpler operation usually translates to a smaller area data-path which in turn translates to more versions of the data-path replicated on the chip. DIEP is optimized for speed, so FPGA area utilization is secondary to speed.

FPGA software development tools such as the Xilinx System Generator, and the Matlab Simulink blocks exist to automatically produce the net-list to place and route into the FPGA without having the write extensive VHDL code. In-circuit evaluation tools and power and thermal analysis tools automate much of the test and verification process.

5. Summary

A DIEP hardware platform using FPGAs will easily keep pace with high speed OCR to speed up and enhance MT using a software architecture composed of image analysis and understand, and adaptive enhancement algorithms. The effect of a properly pipelined hardware solution will be to increase processing time by three orders of magnitude. Near real-time processing of multiple degraded documents enhancements is possible using FPGA technology: 2 months->29 minutes!

Many types of degradation processing can be automatically run in a sequence that "does-no-harm" if the degradation does not exist. Process time is not significant using hardware. DIEP can be ported to create specialized enhancement processing for any specific degraded document types, languages, and applications.