Document Image Quality Assessment for Manual/Automatic Data Extraction

Project Title: Document Image Quality Assessment for Manual/Automatic Data Extraction

Project Number: IMURA0456

Monash Supervisor(s): Prof. Wai Ho Li
Wai.Ho.Li@monash.edu

Monash Primary Contact: Jamie Evans
jamie.evans@monash.edu

Monash Head of Department: Electrical & Computer Systems Engg

Monash ADRT: Emmanuelle Viterbo

IITB Supervisor(s): Prof. Ganesh Ramakrishnan

IITB Primary Contact: ganesh@cse.iitb.ac.in

IITB Head of Department: S Sudarshan
sudarsha@cse.iitb.ac.in

IITB Department: Computer Science and Engineering

Research Academy Themes:
Highlight which of the Academy’s Theme(s) this project will address?
(Feel free to nominate more than one. For more information, see www.iitbmonash.org)
1. Advanced computational engineering, simulation and manufacture
2. Infrastructure Engineering
3. Clean Energy
4. Water
5. Nanotechnology
6. Biotechnology and Stem Cell Research

The research problem
Data entry from physical forms like order forms and invoices is an essential exercise for digitization of data in business process outsourcing. Digitization of essential data from such semi-structured or completely unstructured forms are usually performed manually. Some software tools enable automation of data entry exercise. Automatically extracted data from these software are manually verified and corrected if necessary. Here, automated extraction is achieved through the use of well-defined...
templates over an Optical Character Recognition (OCR). Each template contains a set of general constructs/rules for recognition of textual data using optical character recognition engine and mapping recognized data to essential fields for extraction. Templateization also fails to extract data due to change in expected quality of the document. This change in quality/degradation could occur at various stages of the life-cycle of a document and degradation often lead to difficulties for image processing/analysis, optical character recognition and data extraction at subsequent stages. Degradation of the document could affect document readability also affecting manual data entry. Therefore, it is important to analyse various reasons for loss of document image quality and develop effective methods for automatically measuring these levels of degradation. Poor document images can be rescanned with specified quality metrics for a document image type.

**Project aims**

**Define the aims of the project**

The main aim of the project is to develop techniques for (i) automatically detect document image degradation both local as well as global for manual and automatic data extracted, (ii) classify degradation that lead to poor readability and poor OCR, (iii) perform survey to identify or develop computational metrics for measuring various degradation types (iv) identify gaps in quality metrics and their implementation on various document types and (v) propose an effective approach for quantifying document quality for known and unknown document types.

**Expected outcomes**

**Highlight the expected outcomes of the project**

Create document image database for various manual and automated data extraction tasks
Quantify various data degradation for manual and automated data extraction
Develop rule based, statistical and/or machine learning models for measuring document image quality
Recommend scan parameters for various document types based on observation from the content poor quality image for manual or automatic processing of the document (content aware scanning)

**How will the project address the Goals of the above Themes?**

**Describe how the project will address the goals of one or more of the 6 Themes listed above.**

Data entry from physical forms like order forms and invoices is an essential exercise for digitization of data in business process outsourcing. Digitization of essential data from such semi-structured forms are utmost important to improve productivity and reduce manual efforts. Poor document image quality either due to local or global degradation is a major concern for delaying document processing. While guideline can be generated for digitization of document but they are not enough to rule out several types of degradation as a result of document content. Poor document images could be sent for rescan, part of the poor quality document image could be automated or the entire document could be processed manually. Automatically measuring quality for manual and automatic data entry for various document types and degradation will be useful in efficiently automating the scanning and processing of document images in business process outsourcing.

**Capabilities and Degrees Required**

List the ideal set of capabilities that a student should have for this project. Feel free to be as specific or as general as you like. These capabilities will be input into the online application form and students who opt for this project will be required to show that they can demonstrate these capabilities.

The student should be familiar in statistics and data mining/machine learning algorithms. It will be an advantage if the student is good at text analytics and Image processing.

**For Industry Partners::**

**Potential Collaborators**

Please visit the IITB website www.iitb.ac.in and Monash Website www.monash.edu to highlight some potential collaboratoes that would be best suited for the area of research you are intending to float.

Prof. Ajit Rajwade, IIT Mumbai