IITB-Monash Research Academy

An Indian-Australian research partnership

Indian Institute of Technology Bombay | MONASH University

Project Title:

Learning System for Document Imaging: Data and Pattern Discovery from OCR Extracted Data

Project Number **IMURA0455**

Infosys POWERED BY INTELLECT DRIVEN BY VALUES

Monash Supervisor(s)	Prof.Mark James Carman	*Full names and titles*
Monash Primary Contact:	mark.carman@monash.edu	*Email, phone*
Monash Head of Department:	Judithe Sheard Judy.Sheard@monash.edu	*Full name, email*
Monash Department:	Caulfield School of IT	*Full name*
Monash ADRT:	Sue Mclemish sue.mckemmish@monash.edu	*Full name, email*
IITB Supervisor(s)	Prof.Pushpak Bhattacharya pb@cse.iitb.ac.in Ganesh Ramakrishnan ganesh@cse.iitb.ac.in	*Full names and titles*
IITB Primary Contact:		*Email, phone*
IITB Head of Department:	S Sudarshan sudarsha@cse.iitb.ac.in	*Name, Email,*
IITB Department:	Computer Science and Engineering	*Full name*

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

Organizations daily process large number of documents of different formats and populates the data into databases. Typically, Optical Character Recognition (OCR) software is useful to extract data from scanned images. Automated extraction is achieved through the use of well-defined templates. The data is mostly composed of alphanumeric and other characters, and templates are created to arrive the structure of image document for extraction of data fields, and the correctness of data is determined based on regular expressions. However, due to uncertainty involved in the document content representation, the extracted data sometime results into noise, and hence the accuracy of OCR extraction is limited. Reconstructing the original text for such noisy data is a challenging task. Moreover, creating an appropriate template and updating it based on the previous corrections is tedious. This project involves building a knowledge base for document imaging and discovering meaningful patterns from OCR extracted data in order to improve the accuracy of the field values. The techniques include soft matches, n-gram models, Heuristics, approximation models, etc. The extracted patterns facilitate building accurate reconstructing of text. Further, in addition to data fields, the document may contain the signature (ex. loan documents) which needs to be verified. This project also aims at verifying handwritten signatures.

Project aims

Define the aims of the project

The main aim of the project is to build a learning system that

(i) automatically detects and correct OCR errors
(ii) Predict the accurate threshold values for each variable of interest,
(iii) Verification of handwriting signatures in order to improve the efficacy of the decision-making capabilities of the system.
(iv) Extracting text value occurring once or several times in the document in the form of list or table.
(v) Recognition of handwritten text from structured forms

Expected outcomes

Highlight the expected outcomes of the project

Create a Knowledgebase
Development of statistical and machine learning models for learning system
Development of handwritten signature verification module
Development of Scalable and Improved Document Processing system

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. We use opensource OCR to extract data and the resulting XML files are processed and analyzed to build the knowledgebase using computational intelligence approaches. These computation models results into improved business processes for organizations.

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 collabortoes that would be best suited for the area of research you are intending to float.