Mini Project for IVP course - *Tabular Data Extraction from Document Images*

The project will help you develop a method which can be used to address a general problem of digitizing/extracting tabular data from images of documents/papers containing tabular data with several applications. Paper forms are widely-used for data collection in various surveys, records (ex. patient centric medical records). But accessing and analysing data collected on paper is difficult. Manual data entry is time-consuming and error-prone. A large amount of documents produced in an enterprise also contain tabular data. Such tables convey relationship between multi-dimensional data elements and attributes in a compact (high information density) and easily readable format. For example, in an invoice document, the line items are placed in a tabular from which conveys the number of items purchased, amount etc.

To make the problem more specific, here we will focus on a specific use case for a teaching faculty. Each course is designed with specific Course Outcome's (CO's) and the attainment of CO's can be quantified with the help of assessment techniques. The quantification methodology involves detailed question paper articulation wherein each question is mapped to a CO. Based on marks scored by each student for various questions, a percentage attainment for various CO's can be calculated by the faculty for his course. This could serve an important feedback to the faculty to adjust the course content and/or delivery to improve the attainment of the outcome in future. The problem, however, again is the time consuming and effortful manual process of entering marks scored by every student for each question in all the exams/evaluations. One possible solution that can be implemented is as follows – there is a summary table of marks scored by student on the first page of every answer script which is usually filled by the faculty during the correction and totalling process. After this process, photos of the top page (or only the tabular portion) of the answer scripts can be quickly and easily taken by a normal phone camera. These images can then be analysed by a software to extract the marks from the table automatically thus avoiding the process of manual data entry.

The objective of this project is to develop a program/software using image processing and computer vision techniques to extract accurate data from tables present in these images. You need to exploit the structural properties and other information/constrains relating to the text region within the document. You may make certain assumptions about the input images (regarding illumination condition, orientation, noise, clutter etc.) to begin solving the problem but then try to relax them to get a more general solution that is applicable for wider scenarios. To help you begin your development and testing, we have included some sample images in the folder.

The evaluation of this project will be done in two stages. The expected output of first stage is to be able to extract the image regions corresponding to each table cell that contains a manual entry of marks (along with the identification of the corresponding Q no./part). The deadline for this stage is 30th September. In the second stage your solution should be able to relax any restrictive assumptions made (if any) and recognize the marks accurately and thus able to generate a digital record of information extracted from all the input images. Evaluation of the final output will be done after the second sessional exams.