

Recognition of Forms and Tabular Data

Christos Papadopoulos, Dr Apostolos Antonacopoulos
Department of Computer Science

This project is related to extracting and representing information contained on images of forms and tables, thus allowing for better indexing and archiving of the information available on documents of that type.

As in many Document Image Analysis systems, some of the problems that are addressed in this application include skew detection and correction, noise removal and image quality improvement of the scanned documents. For this project though, the most challenging problems arise with issues that are specific to the subject of forms. Forms are documents that contain structured information around a specific subject. There are numerous different types of form designs and layouts that need to be considered in order to create a system that can automatically recognise and archive forms. This is basically due to the number of different layout design methods that have been used to create forms.

Some of the most important factors that need to be considered when working with such documents include: the method used to segment the form into parts for different data that is printed on it —this could be done either by pre-printed black lines or by white space—, whether the data on the form is positioned in pre-defined spaces or not and whether there is pre-printed information on the form or not.

Currently there are a number of systems that process forms, but for the vast majority of forms it is almost impossible for any of those systems to work flawlessly without any user intervention. One way that many systems achieve acceptable results is by limiting the number of different forms they can process. This limitation might be a successful approach to achieving a good enough success rate for the application, but apart from the obvious drawback of the limitation imposed, it causes problems in other areas such as compatibility with other applications.

This leads to the conclusion that one area that still needs improvement is the representation of forms. Currently most of the representation methods and algorithms are concentrated around the specific purpose of the application in which they are used. In other words, even though they might be perfect for the specific application for which they are developed, they are not generalised to be used to represent any type of form. This lack of a uniform representation that can be used in every form processing application causes a number of problems including incompatibilities between different applications.

A major objective of this project is to implement a form representation method that can be used for most types of forms available; without causing any overhead to an application that processes a specific type of forms, therefore will not use the full functionality of the representation method. Having a representation method that covers most form documents will be ideal, since then the research can be concentrated in further improving other areas of form recognition instead of devising a representation algorithm for every type of form an application might need to deal with.