Abstract

With the recent explosive growth in the availability of digital content, it has become increasingly necessary to develop automated methods of extracting and retrieving relevant pieces of information and presenting them in some readily usable form. As a result of this, the related fields of *information extraction* and *document understanding* have received much attention as of late. One of the sub-areas of these fields, and the focus of this thesis, is the area of *table recognition and extraction*, also referred to as *table understanding*. As tables are widely used to compactly represent information, *table recognition and extraction* has become the focus of much recent research and a wide array of academic and commercial systems have been developed around this research. The goal of these systems is largely to *recognize* and *extract* tables from documents of various formats; the most common formats being image, plain-text, HTML, and PDF. In this thesis we focus on the problem of *pdf table recognition and extraction*, as PDF has become the de facto standard for document sharing and is currently one of the most widely used formats for information exchange.