Classification of Scanned Documents
Master Thesis

Contact

Robert Szabo
Medius AB
+46733378652
robert.szabo@medius.com

Introduction

More and more documents are being generated electronically but there are still a lot of companies relying in manual paper centric processes. It is both from an efficiency and environmental perspective important to drive the invoice processes to more automation. By identifying the manual paper centric processes, Medius could drive the development towards more automation and advise its customers to switch to electronic processes, which in turn has several advantages from an efficiency and environmental perspective.

Medius is a company working in the purchase to pay and document automation area and especially with invoice workflows. With many years of experience around the data capture area, as part of the invoice workflow, we are working with different machine learning techniques to extract information from scanned and electronic documents and to automate the workflows.

Problem formulation

One big part of the invoices that Medius processes are PDF files. These PDF files has either been generated electronically by a financial system or been scanned from paper. The PDF format has different "layers" so it could contain image information, textual information, and algorithms for glyph/drawing generation. Even though the paper invoices have been scanned, many times the scanners themselves contains OCR functionality and produces textual information. Which mean that it is not so easy to accurately determine if a PDF has been scanned or not solely by finding the text layer.

The problem to solve is to find a technique/algorithm that could determine if the PDF has been scanned from paper or if it has been electronically generated. The accuracy of the classification step should be above 99% and it should be fast (below 0.1 second per page).

During the master thesis period, Medius will provide material that could be analyzed and be used for different types of algorithms.