Abstract

Document Layout Analysis, applied to handwritten documents, aims to automatically obtain the intrinsic structure of a document. Its development as a research field spans from the character segmentation systems developed in the early 1960s to the complex systems designed nowadays, where the goal is to analyze high-level structures (lines of text, paragraphs, tables, etc) and the relationship between them.

This thesis first defines the goal of Document Layout Analysis from a probabilistic perspective. Then, the complexity of the problem is reduced, to be handled by modern computing resources, into a set of well-known complementary subproblems. More precisely, three of the main subproblems of Document Layout Analysis are addressed following a probabilistic formulation, namely Baseline Detection, Region Segmentation and Reading Order Determination.

One of the main contributions of this thesis is the formalization of Baseline Detection and Region Segmentation problems under a probabilistic framework, where both problems can be handled separately or in an integrated way by the proposed models. The latter approach is proven to be very useful to handle large document collections under restricted computing resources.

Later, the Reading Order Determination subproblem is addressed. It is one of the most important, yet underestimated, subproblem of Document Layout Analysis, since it is the bridge that allows us to convert the data extracted from Automatic Text Recognition systems into useful information. Therefore, Reading Order Determination is addressed and formalized as a pairwise probabilistic sorting problem. Moreover, we propose two different decoding algorithms that reduce the computational complexity of the problem.

Furthermore, different statistical models are used to represent the probability distribution over the structure of the documents. These models, based on Artificial Neural Networks (from a simple Multilayer Perceptron to complex Convolutional and Region Proposal Networks), are estimated from training data using supervised Machine Learning algorithms.

Finally, all the contributions are experimentally evaluated, not only on standard academic benchmarks but also in collections of thousands of images. We consider handwritten text documents and handwritten musical documents as they represent the majority of documents in libraries and archives. The results show that the proposed methods are very accurate and versatile in a very wide range of handwritten documents.