

Master Thesis on Automatic Information Recognition and Extraction based on Document Image Processing

The focus will be on the automation of information extraction in libraries based on Document Image Segmentation and Recognition. While artificial intelligence (AI) technologies, especially deep learning (DL) systems, have demonstrated state-of-the-art performance especially in the area of natural image processing (e.g. image classification, image segmentation and image generation), a similar break-through has yet to be achieved in the area of document image processing. In particular, document images are notably different from their counterparts, where an accurate comprehension of a document requires not only the recognition of the primitive shapes (e.g. letters or lines which are often extracted as connected components) contained in the image of the document, but also the different homogeneous regions (e.g. a title line or a paragraph) that consist of these basic shapes and, to make things more challenging, the interaction between the regions. This requires development of new AI frameworks capable of hierarchical reasoning across different layers, which will be investigated in the context of this thesis. In the meanwhile, this project is also motivated by the idea of linking research to industry, where the student will translate relevant scientific discoveries to applications that can be implemented in libraries such that key information (e.g. title, author, etc.) of different publications can be automatically recognized and extracted based on the processing of the images of relevant pages of the publications.

This project is jointly promoted by VUB and KBR (Royal Library of Belgium), where the prospective student will have the chance to connect to the cultural heritage sector. Students with background in mathematics, computer science/engineering, electronic and electrical engineering are welcomed.

Interested? Send an email to Tan.Lu@vub.be and Ann.Dooms@vub.be