# Universidade de Pernambuco
# Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

## Proposta de Tese de Doutorado

**Área: Inteligência Computacional / Processamento Digital de Imagens**

**Título: Investigação e proposta de modelo de detecção, reconhecimento e extração de texto estruturado em documentos de layout complexo**

**Orientador – Byron Leite Dantas Bezerra (byronleite@ecomp.poli.br)**

**Coorientador –**

**Descrição**

Em sistemas OCR (*Optical Character Recognition*), a análise de layout e o reconhecimento de texto são de fundamental importância entre as etapas da extração de conteúdo automatizada de uma imagem de documento. Nesse contexto, o reconhecimento de documentos com layout simples pode ser considerado resolvido, enquanto os complexos, como os encontrados em revistas e alguns artigos técnicos, ainda representam um grande desafio [1].

Nos últimos anos, o desenvolvimento de algoritmos ponta a ponta para detecção e reconhecimento de texto a partir de imagens tem evoluído significativamente devido a utilização de modelos de aprendizado profundo (*Deep Learning*). Da mesma forma, a utilização de modelos de Inteligência Artificial em Processamento de Linguagem Natural (*Natural Language Processing*) possibilitou avanço na extração de informações a partir de texto estruturado. Com a capacidade aprimorada desses modelos, foi possível substituir, por exemplo, os métodos clássicos de detecção, reconhecimento do texto e extração de informação de interesse [2,3].

No entanto, além da complexidade da escrita cursiva, a grande diversidade dos documentos também oferece desafios à precisão de sistemas HTR (*Handwritten Text Recognition*), principalmente com layout complexo, como documentos históricos e comerciais, como notas fiscais e cheques bancários. Além disso, sistemas HTR precisam manter a autonomia e robustez no processo de detecção e reconhecimento do conteúdo principal, descartando informações irrelevantes da imagem [4, 5].

De modo geral, em sistemas OCR e HTR são implementados os seguintes passos: (1) captura da imagem do documento; (2) pré-processamento da imagem; (3) análise de layout; (4) seleção e recorte do campo de interesse; (5) pré-processamento da imagem selecionada; (6) reconhecimento do conteúdo, que pode vir na forma impressa ou manuscrita [6].

Dessa forma, o escopo deste projeto de doutorado compreende do 2º até o 6º passo e tem como objetivo a investigação e desenvolvimento de um modelo de aprendizado profundo ponta a ponta no processo de detecção, reconhecimento e extração de texto estruturado em documentos de layout complexo.

**Referências Bibliográficas**

1. A. Antonacopoulos, D. Bridson, C. Papadopoulos, S. Pletschacher. A Realistic Dataset for Performance Evaluation of Document Layout Analysis. Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR2009), Barcelona, Spain, July 2009, pp. 296-300.
2. Long, S., He, X., & Yao, C. (2020). Scene Text Detection and Recognition: The Deep

Learning Era. International Journal of Computer Vision.
3. Minh-Tien Nguyen, Dung Tien Le, Linh Le. Transformers-based information extraction with limited data for domain-specific business documents. Engineering Applications of Artificial Intelligence, v. 97, 2021.
4. J. Puigcerver. Are multidimensional recurrent layers really necessary for handwritten text recognition? 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 67–72, 11 2017.
5. J. A. Sánchez, V. Romero, A. H. Toselli, M. Villegas, E. Vidal. A set of benchmarks for handwritten text recognition on historical documents. Pattern Recognition, v. 94, p. 122–134, 2019.
6. Byron L. D. Bezerra, Cleber Zanchettin, Alejandro H. Toselli, and Giuseppe Pirlo (Eds.). Handwritten: Recognition, Development and Analysis. New York: Nova Science Publishers, 2017.