

Universidade de Pernambuco

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

Proposta de Dissertação de Mestrado

Área: Computação Inteligente / Processamento Digital de Imagens

Título: Aplicação e desenvolvimento de técnicas de Processamento de Linguagem Natural para aprimoramento de sistemas de reconhecimento de escrita

Orientador – Byron Leite Dantas Bezerra (byronleite@ecomppoli.br)

Coorientador – Alejandro H. Toselli (ahector@prhlt.upv.es)

Descrição

Cheques bancários, notas fiscais e fichas de cadastramento são alguns dos formulários que requerem grande esforço manual no processo de digitação de seus dados em sistemas de informação. Como forma de automatizar este processo, é cada vez mais comum o uso de sistemas de processamento automático de documentos em conjunto com tecnologias de OCR (*Optical Character Recognition*) [1].

De um modo geral, são implementados os seguintes passos nestes sistemas: (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 (com letra de forma ou cursiva).

O escopo deste projeto de mestrado reduz-se ao 6º passo, no qual tradicionalmente são usadas técnicas baseadas no *Modelo de Markov Escondido* (HMM) [2] e em Redes Neurais Artificiais (RNA) [3]. Recentemente, arquiteturas mais complexas de RNAs, popularmente conhecidas com o termo *Deep Learning*, têm sido empregadas e alcançado resultados promissores [4,5]. Todavia, quando o vocabulário é irrestrito ou quando o reconhecimento é independente de escritor, os sistemas atuais não atingem resultados satisfatórios para diversos cenários de aplicações demandadas na atualidade [1]. Nesse sentido, o presente projeto de mestrado tem como propósito investigar a aplicação de técnicas de Processamento de Linguagem Natural para aprimoramento dos sistemas de reconhecimento de escrita.

A proposta está relacionada ao projeto “*Soluções de Reconhecimento de Escrita e Processamento de Imagens para BPO de Documentos*”, fomentado pelo CNPQ a partir da Bolsa de Produtividade em Desenvolvimento Tecnológico e Industrial (DTI) deste orientador de mestrado.

Este projeto prevê a colaboração com a iniciativa privada (*Document Solutions*), com o grupo de pesquisa “*Reconhecimento de Padrões*” (RPPDI) da Universidade de Pernambuco, e com o Prof. Dr. Alejandro H. Toselli, pesquisador do grupo “*Pattern Recognition and Human Language Technology Center*” (PRHLT), da Universitat Politècnica de València.

Neste projeto é requerida a dedicação exclusiva do aluno, visto que será oferecida uma bolsa de mestrado mediante parceria com a *Document Solutions*.

Referências Bibliográficas

1. Byron L. D. Bezerra, Cleber Zanchettin, Alejandro H. Toselli, and Giuseppe Pirlo (Eds.). *Handwritten: Recognition, Development and Analysis*. New York: Nova Science Publishers, 2017.
2. A. H. Toselli and E. Vidal, “Handwritten text recognition results on the Bentham collection with improved classical N-gram-HMM methods”, in *Int. Workshop on Historical Document Imaging and Processing (HIP)*, Nancy, France, August 2015.
3. S. España-Boquera, M.J. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martínez, “Improving Off-line Handwriting Text Recognition with Hybrid HMM/ANN Models”. *IEEE Transactions on PAMI*, vol. 33, no. 4, pp. 767–779, 2011.
4. J. Andreu Sanchez, V. Romero, A. Toselli, and E. Vidal, “ICFHR2014 Competition on Handwritten Text Recognition on Transcriptorium Datasets (HTRtS)”, in *14th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR)*, Sep. 2014, pp. 785-790.
5. Diem et al. (2014), *ICFHR 2014 Competition on Handwritten Digit String Recognition in Challenging Datasets (HDSRC 2014)*, ICFHR 2014.