Instituto de Ciências Matemáticas e de Computação | Universidade de São Paulo |

Extração e Análise pelo Contorno de Palavras de Textos Históricos

Lucas Schmidt Cavalcante Gustavo Enrique de Almeida Prado Alves Batista João do E. S. Batista Neto
Instituto de Ciências Matemáticas e Computação - ICMC

Motivação

Existe atualmente um enorme esforço para digitalizar grandes acervos de documentos históricos. A existência de uma imagem digitalizada auxilia a perpetuar a memória do documento, e tal imagem pode ser utilizada em futuras pesquisas. Entretanto, textos disponiblilizados na forma de imagens limitam a capacidade de pesquisar e correlacionar as informações presentes neles. É preciso que haja um modo de fazer busca, e para isto é preciso transcrever esses textos, transformando-os de imagens para cadeias de caracteres. Como reconhecer escrita cursiva ainda é um problema em aberto, opta-se por atacar um problema semelhante, que é a identificação da ocorrência de todas as instâncias de uma mesma palavra, com isto um humano só precisa transcrever uma instância de cada palavra, reduzindo assim o trabalho de transcrição a uma fração do total [1].

Metodologia

O método se resume a medir a semelhança de contornos de palavras utilizando a medida *Dynamic Time Warping*. Por meio dessa distância é realizado um agrupamento dos contornos, sendo que contornos (e portanto palavras) semelhantes são mapeadas para os mesmos *clusters*. Resta a um analista humano identificar e transcrever as palavras agrupadas em um *cluster*. O método proposto envolve técnicas de processamento de imagens e inteligência artificial. Na parte de processamento de imagem é feita a segmentação, aplicação do processo de *skeletonizing* e de *prunning*, tratamento de ruído e determinação das palavras. Na parte de inteligência artificial são extraídas as séries "temporais"(contornos) de todas as palavras para a medição de distância entre elas. Todo o processo pode ser visto no diagrama abaixo.

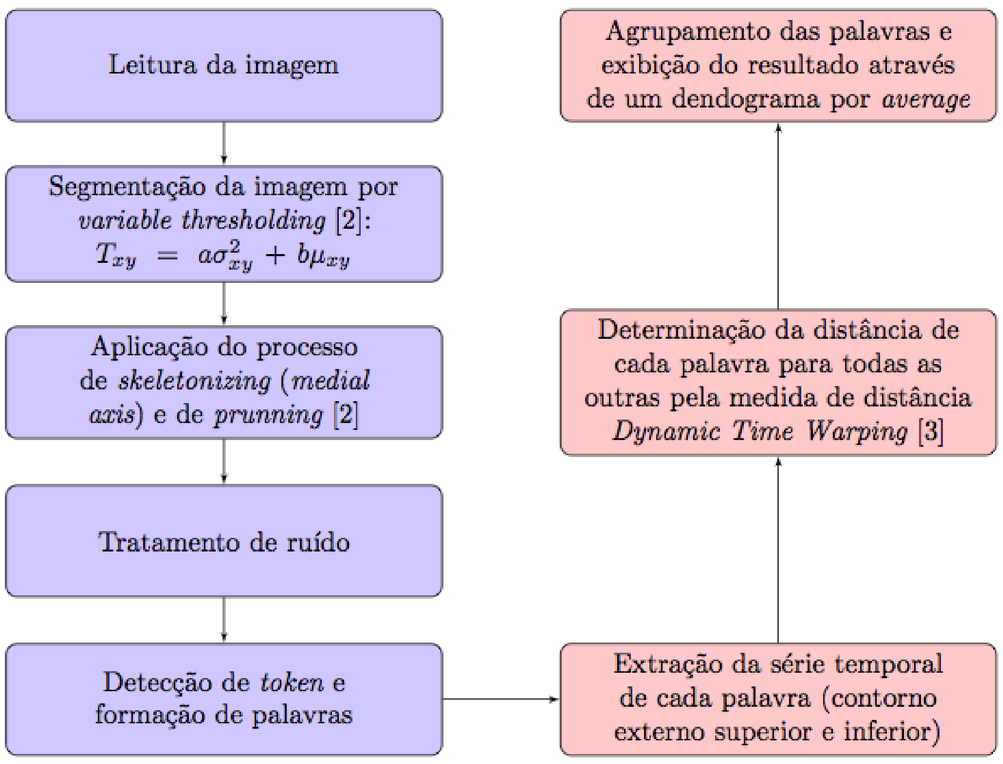

Figura: Diagrama para o processo de obtenção do agrupamento das palavras.

Resultados

A figura abaixo mostra o resultado da aplicação das técnicas de processamento de imagem em um documento histórico.

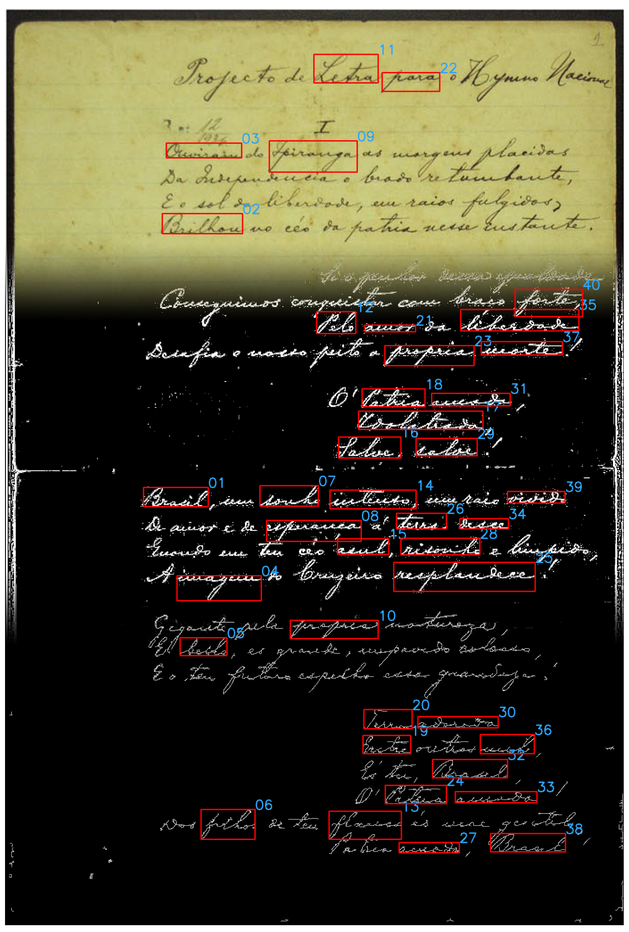

Figura: Ilustração da etapa de processamento de imagem em um documento histórico de escrita cursiva retirado da Biblioteca Nacional Digital [4].

O agrupamento das palavras do documento acima pode ser visto no dendograma (por *average linkage*) abaixo.

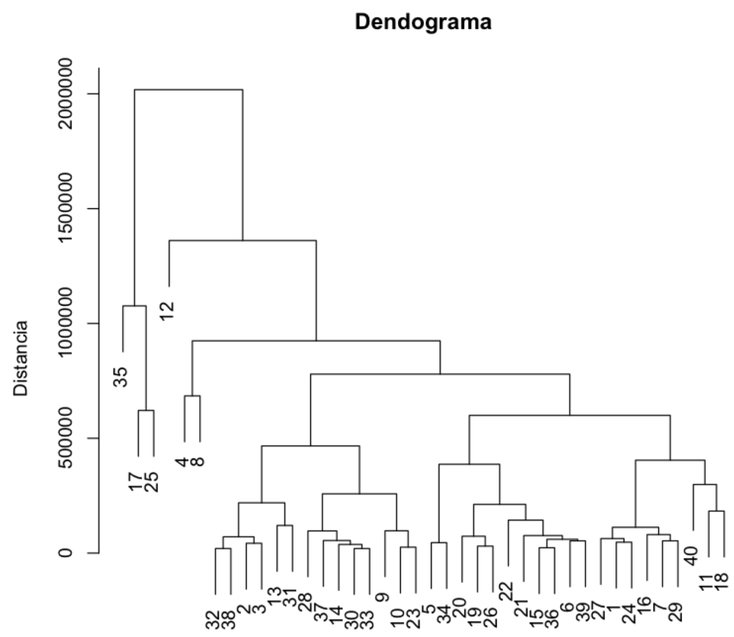

Figura: Agrupamento das palavras do documento histórico.

Conclusão

Com os resultados obtidos confirma-se o indicativo de viabilidade de uso de séries temporais para fazer o agrupamento de palavras extraídas de imagens de documentos históricos de escrita cursiva.

Referências

[1] Rath, T. M. and Manmatha, R. (2003). Word Image Matching Using Dynamic Time Warping. Computer Vision and Pattern Recognition, IEEE Computer SocietyConferenceon, 2:521.
[2] Gonzalez, R. C. and Woods, R. E. (2007). Digital Image Processing. PrenticeHall, 3edition.
[3] Keogh, E. and Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. Knowledge and Information Systems, 7(3).
[4] Biblioteca Nacional Digital: http://bndigital.bn.br/.

Apoio

Conselho Nacional de Desenvolvimento Científico e Tecnológico