XIV Computer on the Beach
30 de Março a 01 de Abril de 2023, Florianópolis, SC, Brasil

Zeig Dich: Dataset para Reconhecimento de Tipos de Fonte de Jornais Históricos Teuto-Brasileiros

Lucas Sulzbach
Universidade Federal do Paraná
Curitiba, Paraná, Brasil
lucas.sulzbach@ufpr.br

Thomas Bianchi Todt
Universidade Federal do Paraná
Curitiba, Paraná, Brasil
thomas.todt@ufpr.br

Pedro Domingos Tricossi dos Santos
Universidade Federal do Paraná
Curitiba, Paraná, Brasil
pedro.tricossi@ufpr.br

Thalita Maria do Nascimento
Universidade Federal do Paraná
Curitiba, Paraná, Brasil
thalitanascimento@ufpr.br

Eduardo Todt
Universidade Federal do Paraná
Curitiba, Paraná, Brasil
todt@ufpr.br

ABSTRACT

This paper addresses the challenge of typeface recognition, within the broader scope of optical character recognition of historical German-Brazilian periodicals. A dataset of words containing annotations of font types and transcriptions for training neural networks for typeface and text recognition is presented. By enabling word-level typeface and text recognition, the authors plan to later develop techniques for high-precision OCR of historical prints typeset in heterogeneous font styles. The value of this dataset is proven by the excellent results obtained by artificial neural networks trained on it. The authors also recognize that even better results can be obtained by exploring new ways of organizing the dataset prior to training, and that the results can also be improved through modifications in the architecture of the nets used.

PALAVRAS-CHAVE

Datasets, Redes Neurais, OCR, Documentos Históricos, Humanidades Digitais

1 INTRODUÇÃO

Durante os séculos XIX e XX, o Brasil recebeu milhares de imigrantes de origem alemã. Esses imigrantes foram responsáveis pela fundação de diversas colônias teuto-brasileiras ao redor do país, aonde o idioma alemão foi preservado por gerações. Nestas comunidades, o idioma se fez presente não apenas na tradição oral, mas também na escrita, através de centenas de publicações periódicas como jornais, revistas e almanaques em língua alemã que circularam ao longo das décadas. Nos dias atuais, exemplares de documentos produzidos por essa imprensa teuto-brasileira possuem grande valor histórico e linguístico, e podem ser encontrados em acervos, bibliotecas e coleções pessoais.

Com os avanços tecnológicos, a digitalização de documentos antigos vem ganhando força ao redor do mundo como estratégia de preservação e facilitação do acesso ao patrimônio histórico e cultural. Neste contexto surge o projeto *dbp digital*[1], responsável por escanear alguns dos acervos de periódicos teuto-brasileiros, dando origem a um *corpus* digital da imprensa brasileira em língua alemã. A digitalização, porém, vai muito além do simples ato de escanear, isto é, de capturar imagens das páginas de documentos, pois também deve tratar de transcrever os textos destas páginas escaneadas e produzir documentos digitais pesquisáveis, que permitam a um pesquisador fazer buscas pelo conteúdo de uma página, um documento, ou mesmo de toda uma coleção de documentos.

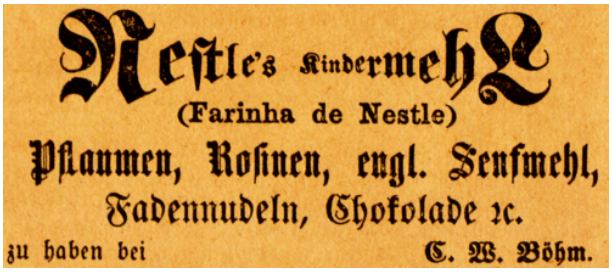
Para implementar um processo de transcrição de imagens em larga escala, se faz necessário o emprego de ferramentas de *reconhecimento óptico de caracteres* (OCR – *Optical Character Recognition*), que, em suas versões mais recentes, fazem o uso de redes neurais treinadas para reconhecer textos contidos em imagens. Infelizmente, as principais ferramentas de reconhecimento de texto se mostram inadequadas e insuficientes para transcrever documentos antigos em língua alemã com precisão [1, 2]. Dentre os motivos para esse problema, é possível apontar a utilização de tipos de fonte antigos na digitação desses documentos, em especial dos tipos góticos, característicos da tipografia alemã da época. No caso dos documentos teuto-brasileiros, existe ainda um agravante, que é o bilinguismo. Na época em que foram impressos, era prática comum utilizar tipos de estilo gótico, como o *textur* e o *fraktur*, para digitar texto em alemão, e empregar fontes de estilo latino, como o *antiqua*, para tipografar texto em outros idiomas. A língua portuguesa, ainda que figurando como um idioma secundário, se faz presente em muitos exemplares de documentos teuto-brasileiros, resultando num alto contraste visual entre tipos góticos e latinos. Além disso, na tentativa de atrair a atenção dos leitores, muitos anunciantes construíam propagandas chamativas com fontes de variados estilos e tamanhos, o que aumenta ainda mais a diversidade de fontes encontradas nesses documentos. Exemplos de anúncios como esses podem ser conferidos na Figura 1. Em razão dessa heterogeneidade tipográfica, é natural que as redes neurais utilizadas por ferramentas de OCR convencionais, treinadas com imagens de documentos de tipografia mais homogênea, não sejam capazes de reconhecer corretamente os textos da imprensa brasileira em língua alemã.

2 SOLUÇÃO PROPOSTA

Tendo em vista a demanda por ferramentas capazes de reconhecer os textos dos periódicos teuto-brasileiros com precisão, é proposto um *dataset* de imagens de palavras tipografadas em diferentes estilos de fonte. Esse *dataset* deve incluir, além das imagens em si, as suas respectivas transcrições e também anotações das classes de fonte correspondentes. O propósito deste conjunto de dados é ser utilizado para treinar dois tipos de redes neurais: um para

[1] https://dokumente.ufpr.br/pt-br/dbpdigital.html

(a) Anúncio de máquina de escrever que escreve em "alemão (*fraktur*), latim (*antiqua*) e cursivo"



(b) Anúncio digitado com tipos de diferentes tamanhos e estilos (*fraktur* (mais de uma variante), *antiqua* e *textur*)

Figura 1: Exemplos de anúncios com fontes heterogêneas

reconhecimento de texto, e outro para classificação deste texto pela fonte utilizada para tipografá-lo. O motivo para a definição dessas duas categorias de redes neurais é que elas deverão ser usadas para experimentar também duas estratégias para reconhecer textos tipografados em múltiplas fontes: A primeira consistindo em utilizar as imagens do *dataset* e suas transcrições para treinar modelos generalistas de reconhecimento de texto, que sejam capazes de reconhecer textos independentemente da fonte em que estejam tipografados. Chamamos este método de *fonte-independente*, e é uma abordagem similar à adotada por Springmann et al. [2018] para treinar redes neurais com textos escritos e tipografados em alemão gótico e em latim. A segunda estratégia, que denominamos *fonte-dependente*, consiste em treinar modelos dos dois tipos: Um meramente para o reconhecimento da fonte utilizada, que utilizará como conjunto de treinamento as imagens e suas respectivas classes de fonte, e um modelo de reconhecimento de texto para cada classe de fonte distinta, treinados apenas com as imagens e transcrições de suas respectivas classes. O raciocínio por trás dessa abordagem é que a rede neural treinada para reconhecer fontes poderá ser usada para identificar a fonte em que o texto está tipografado para então selecionar o modelo de reconhecimento de texto dedicado àquela fonte em específico, que tenderá a ser mais preciso do que o modelo generalista da primeira abordagem. Essa estratégia de classificação de fontes é similar ao trabalho de Seuret et al. [2019], que consiste de um *dataset* de textos medievais em fontes diversas.

A decisão pelo uso de imagens de palavras ocorreu em razão da alta variedade de tipos de fonte comentada anteriormente. Não sendo raras as alternâncias entre os tipos, existem casos em que não é possível atribuir uma única classe de fonte a uma linha de texto, muito menos a um parágrafo ou a uma página inteira. Palavras, por outro lado, muito dificilmente eram tipografadas em mais de um tipo de fonte, o que as torna uma unidade de texto mais adequada para a tarefa de classificação de textos por estilo de fonte. Essa granularidade é uma inovação em relação ao *dataset* apresentado por Seuret et al. [2019], em que as imagens das páginas foram rotuladas pelos tipos de fonte por ordem de frequência.

Devido ao tempo e esforço que as atividades aqui descritas demandam, limitamos o escopo deste resumo estendido à construção do *dataset* e à condução de experimentos de reconhecimento de fontes com redes neurais. O treinamento de modelos de reconhecimento de texto e a implementação e comparação das estratégias *fonte-independente* e *fonte-dependente* deverão ser realizados em trabalhos futuros.

3 RESULTADOS PRELIMINARES

Para a construção do *dataset*, o primeiro passo foi a seleção de uma amostra de imagens de páginas do *corpus dbp digital*. Para garantir a maior diversidade de tipos de fonte possível, foram selecionadas apenas as páginas que continham anúncios. Das mais de 6.000 páginas às quais tivemos acesso, foram selecionadas cerca de 2.000 páginas de jornais com anúncios. Posteriormente, foram selecionados 311 anúncios, visando a composição de uma amostra com alta variedade de fontes. Os anúncios foram segmentados manualmente com a ajuda a ferramenta *LAREX*[2]. Após isso, foi utilizado o *tesseract*[3] para segmentar as palavras dos anúncios de forma automática, já que seria muito dispendioso segmentar manualmente um número de palavras tão alto. Foram então descartadas as imagens em que houveram erros de segmentação, resultando numa amostra de cerca de 10.000 imagens de palavras.

Finalizada a etapa de seleção das imagens, os autores realizaram um estudo para aprender a identificar e a transcrever as fontes encontradas nestas imagens. Inicialmente, as palavras foram separadas em apenas três classes: *Textur*, *fraktur* e *antiqua*. Porém, após uma análise mais aprofundada, foi constatado que algumas fontes latinas classificadas como *antiqua* seriam mais adequadamente classificadas como *italic* (itálico) ou *script* (cursivo). Além disso, os autores julgaram melhor classificar palavras tipografadas em *kanzlei-fraktur* (uma variação de *fraktur*) como uma classe à parte (*kanzlei*), por ser muito distinta das demais variantes de *fraktur*. Exemplos de cada uma dessas classes podem ser conferidas na Figura 2.

Uma vez definidas as classes em que as palavras deveriam ser categorizadas, foi dado início ao processo de classificação e transcrição de cada palavra. Para isso, a ferramenta *neat*[4] foi adaptada, de modo que fosse possível visualizar as imagens do *dataset*, palavra por palavra, e simultaneamente selecionar suas respectivas classes

[2] https://github.com/OCR4all/LAREX
[3] https://github.com/tesseract-ocr/tesseract
[4] https://github.com/qurator-spk/neat

Figura 2: Exemplos das classes de fontes consideradas para o trabalho

e preencher suas respectivas transcrições. Para acelerar o processo, foi utilizado novamente o *tesseract*, desta vez para reconhecer o texto de cada uma das palavras. Apesar das imprecisões, essa etapa se mostrou útil para preencher previamente algumas das palavras e otimizar o tempo durante as transcrições.

Finalizado todo o processo de selecionar, classificar e transcrever as imagens, o *dataset* estava completo. O próximo passo consistiu, então, em utilizá-lo para treinar modelos de reconhecimento de fonte. Nos experimentos iniciais, as redes neurais apresentaram dificuldades ao tentar reconhecer palavras de duas classes de fonte em particular: O *kanzlei*, que acabou por ser sempre classificado como *fraktur*, e o *script*, que foi confundido com o *italic*. Os autores optaram então por desfazer a separação entre *fraktur* e *kanzlei*, agrupando novamente as duas classes, e também por agrupar as classes *italic* e *script*, resultando numa classe única de "variantes de *antiqua*". Além disso, foi feito o uso de operações de *data augmentation* (*image zoom in* e *zoom out, small rotation - 30°* e *hue deviation*). Após estas alterações, os modelos treinados foram capazes de classificar as palavras de forma muito mais precisa.

Para os experimentos, foram utilizadas as arquiteturas *ResNet50v2* e *VGG19*. Os resultados podem ser conferidos na Tabela 1. A matriz de confusão se encontra na Tabela 2.

Arquitetura	AUC	Precision	Recall
ResNet50v2	0.98	0.98	0.97
VGG19	0.99	1.00	0.99

Tabela 1: Resultados para classificação em 4 classes.

	Antiqua	Fraktur+Kanzlei	Italic+Script	Textur
Antiqua	150	3	3	0
Fraktur+Kanzlei	0	838	0	0
Italic+Script	0	0	124	0
Textur	0	0	0	552

Tabela 2: Matriz de Confusão para 4 classes

4 CONSIDERAÇÕES FINAIS

Os resultados apresentados neste trabalho demonstram que a estratégia de classificar textos de periódicos teuto-brasileiros pela fonte utilizada para tipografá-los é promissora. Estudos mais aprofundados são bem-vindos para avaliar se é possível alcançar resultados parecidos ao treinar modelos para reconhecer classes mais específicas de fontes, como por exemplo, as diferentes variantes do *fraktur* (*kanzlei-fraktur*, *fette-fraktur*, etc.) e as fontes itálicas e cursivas que foram agrupadas numa classe só para os fins deste trabalho. Técnicas de pré-processamento de imagens e arquiteturas de redes neurais não exploradas neste trabalho poderão ajudar a atingir esses objetivos.

O *dataset* aqui apresentado servirá também de subsídio para outros trabalhos visando treinar redes neurais para reconhecimento óptico de caracteres. A partir desses novos modelos, poderão ser colocadas a prova os métodos de reconhecimento de texto *fonte-independente* e *fonte-dependente* propostos neste trabalho para assim desenvolver estratégias robustas de reconhecimento de textos não apenas de periódicos teuto-brasileiros, mas de documentos históricos com fontes heterogêneas em geral.

REFERÊNCIAS

[1] Alessandra Belézia Araújo. 2019. Análise de layout de página em jornais históricos germano-brasileiros. https://hdl.handle.net/1884/63706

[2] Clemens Neudecker, Konstantin Baierer, Maria Federbusch, Matthias Boenig, Kay-Michael Würzner, Volker Hartmann, and Elisa Herrmann. 2019. OCR-D: An end-to-end open source OCR framework for historical printed documents. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*. ACM. https://doi.org/10.1145/3322905.3322917

[3] Mathias Seuret, Saskia Limbach, Nikolaus Weichselbaumer, Andreas Maier, and Vincent Christlein. 2019. Dataset of Pages from Early Printed Books with Multiple Font Groups. In *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*. ACM. https://doi.org/10.1145/3352631.3352640

[4] Uwe Springmann, Christian Reul, Stefanie Dipper, and Johannes Baiter. 2018. Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. (2018). https://doi.org/10.48550/ARXIV.1809.05501