Extração de elementos textuais em imagens capturadas por *smartphones*: análise da relação entre as características das imagens e a eficácia da extração

Daniel M. Kuhn[1], Cristiano R. Cervi[1], Edimar Manica[2]

[1]Instituto de Ciências Exatas e Geociências – UPF - Passo Fundo - RS

[2]Campus Ibirubá - IFRS - Ibirubá - RS

138714@upf.br, cervi@upf.br, edimar.manica@ibiruba.ifrs.edu.br

Abstract. *Optical character recognition software is designed to convert textual elements of documents into editable and searchable text. This task presents specific challenges when submitted to images captured by smartphone cameras. This work experimentally analyzes the relationship between extraction efficiency in images captured by smartphones and their characteristics. The experiments demonstrate that images with curvilinear text and light variation do not substantially compromise extraction efficiency, whereas images with inclined texts as well as images with unclear characters have the lowest extraction rates.*

Resumo. *Softwares de reconhecimento óptico de caracteres têm como propósito converter elementos textuais de documentos em texto editável e pesquisável. Essa tarefa apresenta desafios específicos quando submetida a imagens capturadas por câmeras de smartphones. Este trabalho analisa experimentalmente a relação entre a eficácia de extração em imagens capturadas por smartphones e suas características. Os experimentos demonstram que imagens com texto curvilíneo e variação de iluminação não comprometem substancialmente a eficácia de extração, ao instante que imagens com textos inclinados, bem como imagens que possuem caracteres pouco nítidos, apresentam os menores índices de extração.*

1. Introdução

A análise de Big Data é um aspecto chave da sociedade moderna uma vez que permite criar conhecimento a partir de dados. Essa análise traz o conhecimento para o indivíduo de uma forma direta e facilitada permitindo a emancipação das pessoas e as habilitando a agirem e tomarem decisões com mais embasamento [Manica, Dorneles and Galante 2017]. Problemas de heterogeneidade, escalabilidade, complexidade e privacidade impedem o progresso de todos os estágios do *pipeline* que extrai valor a partir de dados [Labrinidis and Jagadish 2012]. Nesse contexto, os problemas iniciam durante a aquisição de dados porque muitos dados não estão nativamente em um formato estruturado e estruturar tal conteúdo para análise futura é o principal desafio [Agrawal et al 2012].

Um exemplo de dados relevantes em um formato não estruturado é observado em textos presentes em imagens postadas nas redes sociais. Estima-se que só no instagram - atualmente a maior rede social de fotografias - são postadas em média 52 milhões de fotografias todos os dias [Statistic Brain, 2017]. A extração de elementos textuais contidos em imagens de trechos de livros postadas na redes social pode ser útil para identificar o que os usuários estão lendo e então recomendar outros livros semelhantes.

A extração de conteúdos textuais em imagens é realizada através do uso de softwares de Reconhecimento Óptico de Caracteres (OCR – *Optical Character Recognition*). O OCR é um processo de reconhecimento visual que converte documentos de texto em texto editável e pesquisável [Berchmans and Kumar 2014]. Nos últimos trinta anos um número substancial de pesquisas acerca de mecanismos de OCR foram realizadas [Islam and Noor 2016]. Em suma, a grande maioria dos esforços destinou-se a solucionar problemas decorrentes da digitalização de documentos de texto através do uso de dispositivos de *scanner*, o que resultou na obtenção de altas taxas de precisão de extração em documentos desta natureza [Asad et al 2016].

Entretanto, os métodos de pré-processamento de imagens aplicados em documentos escaneados são em diversos casos inapropriados ou insuficientes quando destinados a otimizar o reconhecimento de caracteres de imagens capturadas por câmeras de *smartphones*. Isso se deve ao fato de que as características encontradas em imagens escaneadas são, em sua grande maioria, distintas das características presentes em arquivos de imagens obtidas através da câmera de *smartphones*. As imagens capturadas por câmeras podem apresentar baixa resolução, desfocagem e distorção de perspectiva, apresentando layouts complexos e interação entre o conteúdo e o plano de fundo [Liang, Doermann and Li 2005].

Este trabalho tem como objetivo geral avaliar experimentalmente a eficácia da extração de um software de OCR em imagens capturadas por câmeras de *smartphones*. O objetivo específico é relacionar as características das imagens com a eficácia da extração. De acordo com os experimentos realizados, as características que mais impactam a eficácia da extração são: (i) linhas de texto inclinadas; (ii) caracteres pouco nítidos.

Este artigo está organizado da seguinte forma. Na Seção 2, são discutidos os trabalhos relacionados. A Seção 3 descreve a metodologia dos experimentos. Na Seção 4, são discutidos os resultados dos experimentos. Finalmente, a Seção 5 apresenta as considerações finais e os trabalhos futuros.

2. Trabalhos relacionados

O trabalho de [Asad et al 2016] apresenta um sistema de OCR baseado em redes LSTM (*Long Short Term Term*), capaz de reconhecer caracteres borrados decorrentes de movimentos indesejados. Redes LSTM, são um tipo especial de redes neurais recorrentes (RNN – *Recurrent neural network*) com capacidade de recordar informações por longos períodos de tempo [Olah 2015].

Em [Smith 1987] foi proposto uma nova abordagem para reconhecimento de

caracteres. Esse trabalho deu origem ao motor de OCR *Tesseract* [Tesseract 2015]. Em 2005, *Tesseract* passou a ser um projeto *Open Source* e desde 2006 vem sendo desenvolvido pela *Google Inc*. *Tesseract* provê suporte a Unicode, capaz de reconhecer mais de 100 linguagens diferentes [Tesseract 2015]. *Tesseract* é totalmente treinável, sendo possível adicionar novos símbolos e até mesmo novos idiomas inteiros. A possibilidade de aplicar processos de treinamento, bem como, o fato de ser um projeto *Open Source,* foram fatores determinantes para a escolha do *Tesseract* como extrator.

Em [Kuhn, Cervi and Manica 2017] foi avaliada a eficácia da extração de elementos textuais em imagens capturadas por *smartphones* submetidas ao *Tesseract*. Este trabalho diferencia-se por expandir os experimentos e realizar uma análise dos resultados identificando a relação entre as características das imagens e a eficácia da extração.

3. Metodologia

Nesta seção, é descrita a metodologia adotada nos experimentos. A Figura 1 apresenta o fluxo de execução dos experimentos, composto por 6 etapas:

1. **Obtenção** - onde foram coletadas as imagens para compor a base de dados dos experimentos;
2. **Anotação** - onde foram transcritos manualmente os elementos textuais das imagens;
3. **Definição** - onde definiu-se as características de interesse a serem analisadas;
4. **Identificação** - onde identificou-se a presença das características definidas nas imagens da base de dados;
5. **Configuração e execução** - onde submeteu-se a base de dados a um extrator para obter os elementos textuais contidos na imagem de forma automática;
6. **Análise** - onde os resultados obtidos foram analisados, relacionando a eficácia da extração com as características das imagens.

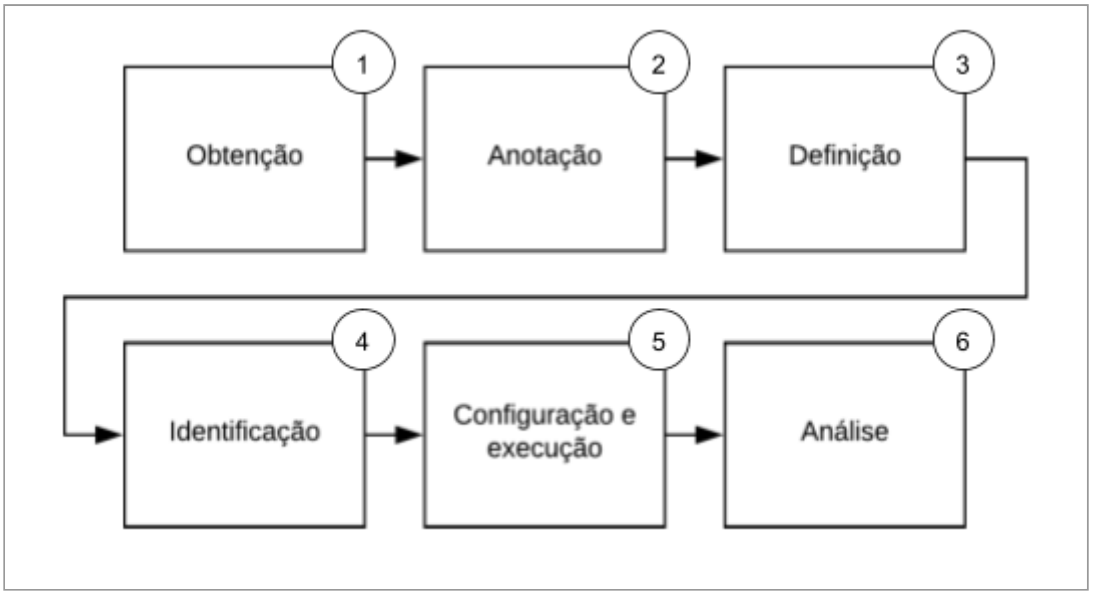

Figura 1. Etapas dos experimentos realizados.

As subseções a seguir detalham as etapas que integram os experimentos.

3.1. Obtenção

Para a obtenção da base de dados, foram coletadas 160 imagens de trechos de livros capturadas por *smartphones*. Desse total, 82 imagens (51.25%) foram coletadas manualmente de publicações de redes sociais. As 78 amostras restantes (48.75%) foram fornecidas por um grupo de voluntários, que possuíam *smartphone* e tinham idade entre 14 e 29 anos.

Os voluntários receberam a tarefa de fotografar pequenos trechos de um livro qualquer utilizando seus *smartphones*. Após, deveriam recortar o trecho, segmentando entre toda a imagem, o trecho de real interesse. Por fim, deveriam enviar a imagem resultante por e-mail ou rede social. Ressalta-se que alguns voluntários não segmentaram a imagem corretamente, ou seja, algumas imagens possuem ruído.

3.2. Anotação

O conteúdo textual contido em cada uma das 160 imagens foi manualmente transcrito por um especialista. O texto em formato digital foi então armazenado na base de dados, mantendo a devida relação entre o arquivo da imagem e seu respectivo conteúdo textual. Dessa forma, gerou-se o gabarito de extração. Todas as imagens estavam legíveis o suficiente para permitir o reconhecimento manual de todas as palavras nelas contidas.

A Figura 2 (a) apresenta um exemplo de imagem da base de dados. A Figura 2 (b) apresenta o gabarito para essa imagem, ou seja, os elementos textuais identificados pelo especialista. Observa-se que essa imagem não foi segmentada corretamente, uma vez que possui caracteres que não pertencem ao texto de interesse.

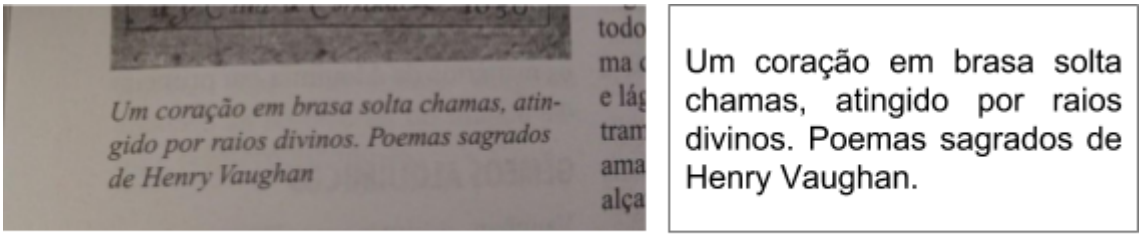

Figura 2: Exemplo de imagem da base de dados e seu respectivo gabarito

3.3. Definição

Nesta etapa, foram selecionadas as características das imagens a serem avaliadas com relação ao seu impacto na eficácia da extração. Foram definidas quatro características: (i) linhas de texto inclinadas; (ii) linhas de texto com aspecto curvilíneo; (iii) variação de iluminação; e (iv) caracteres pouco nítidos. A seguir, cada característica é explicada e exemplificada.

3.3.1. Linhas de texto inclinadas

Esta característica refere-se ao aspecto de inclinação das linhas da imagem. Conforme pode-se observar na Figura 3, a imagem pode estar excessivamente rotacionada no

sentido horário, ou seja, possui graus de inclinação negativos, ou então, pode estar excessivamente rotacionada no sentido anti-horário, e nesse caso, apresenta graus de inclinação positivos.

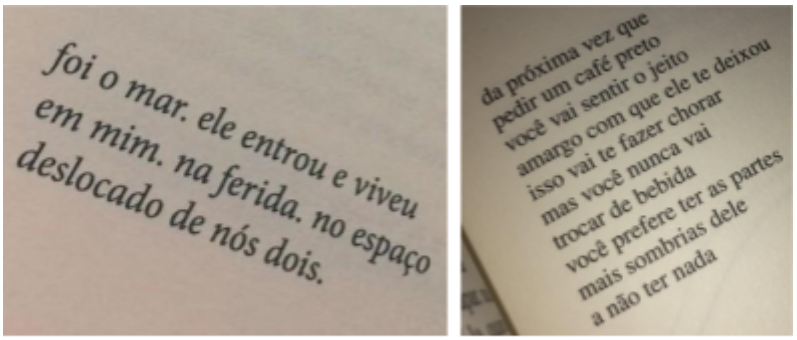

Figura 3: Exemplo de imagens apresentando graus de inclinação: (a) imagem com grau de inclinação negativo, (b) imagem com grau de inclinação positivo.

Essa característica resulta do posicionamento da câmera em relação ao texto a ser fotografado. Ao contrário das imagens escaneadas, onde há a presença de um suporte que sugere a posição correta, as fotografias não possuem posicionamento pré-definido.

3.3.2. Linhas de texto com aspecto curvilíneo

Esta característica é resultado decorrente da perspectiva das páginas em relação à câmera. Visto que as páginas dos livros tendem a curvar-se, quando fotografadas dessa maneira, produzem imagens com distintas perspectivas ao longo da página. Como resultado, as linhas de texto tendem a apresentar aspectos curvilíneos.

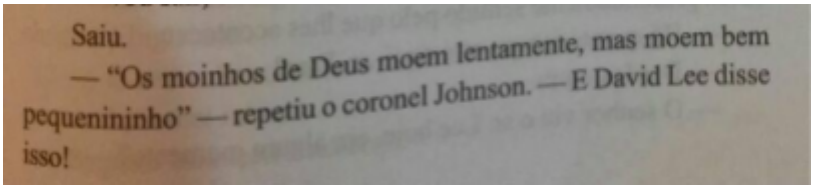

Figura 4. Exemplos de imagens apresentando linhas de texto com aspecto curvilíneo

Pode-se observar na Figura 4 que a imagem apresenta aspecto curvilíneo.

3.3.3. Variação de iluminação

Esta característica refere-se a variações de iluminação sobre a imagem. Nessa categoria, são incluídas também, imagens com presença total ou parcial de sombras.

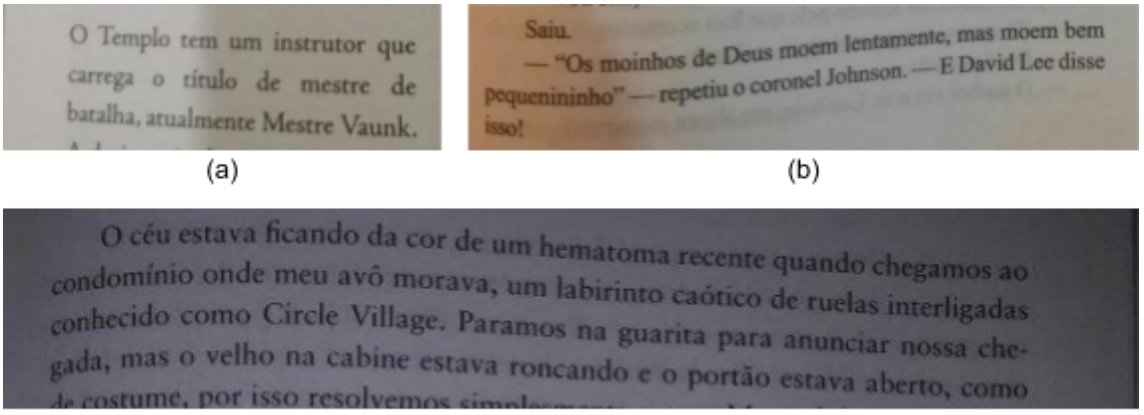

Figura 3. Exemplo de imagens com variação de iluminação

As imagens apresentadas na Figura 3 (a) e 3 (b) apresentam presença parcial de sombra sobre o texto de interesse. A imagem apresentada na Figura 3 (c) apresenta presença de sombra sobre todo o texto de interesse.

3.3.4. Caracteres pouco nítidos

Foram consideradas imagens com caracteres pouco nítidos, aquelas que possuíam caracteres desfocados ou borrados. Essa característica decorre de diversos fatores, entre eles: (i) qualidade da câmera; (ii) foco da imagem; (iii) iluminação - abordado neste trabalho como variação de iluminação; (iv) perspectiva - também abordado neste trabalho com a denominação de linhas de textos com aspecto curvilíneo.

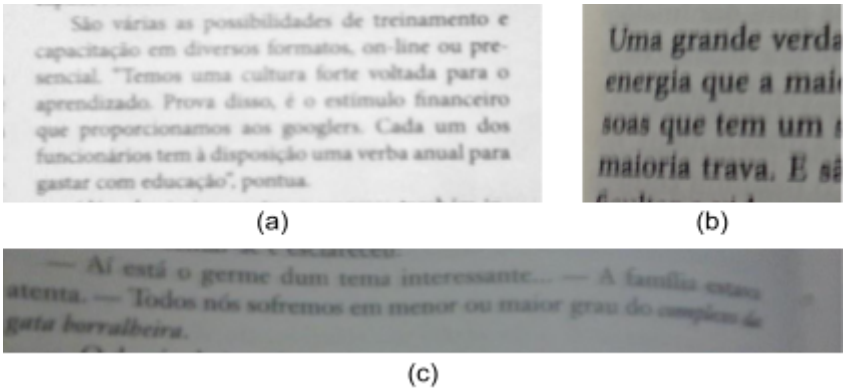

Figura 4: Exemplo de múltiplos cenários relacionados à pouca nitidez dos caracteres

A Figura 4 demonstra imagens contendo diferentes aspectos em que há ocorrência de caracteres pouco nítidos. A Figura 4 (a) foi fotografada com um dispositivo com câmera de baixa qualidade. Os caracteres da Figura 4 (b) apresenta efeito fantasma, decorrente de movimentos indesejados no momento da fotografia. A Figura 4 (c) apresenta uma imagem não focalizada.

4. Identificação

Nesta etapa, identificou-se entre as características de interesse, quais estavam presentes nas imagens da base de dados. A identificação de características nas imagens foi dividida em processos manuais e automatizados. Como processos manuais, com possibilidade binária de resposta (Afirmativo ou Negativo) as seguintes questões foram respondidas por um usuário especialista: (i) as linhas de texto apresentam aspectos curvilíneos?; (ii) é possível identificar variações de iluminação ou presença parcial ou total de sombra sobre a imagem?; (iii) os caracteres do conteúdo textual estão nítidos?. Este processo foi repetido para as 160 imagens.

Para a identificação da inclinação das linhas de texto, utilizou-se um algoritmo[1] capaz de identificar o número (em graus) de inclinação das linhas de texto. O algoritmo utiliza métodos de transformação morfológicas para evidenciar os pixels que formam o segmento da linha de texto e utiliza o método conhecido como Transformada de Hough [OpenCV 2017] para identificar o segmento das linhas de texto com base no número de pontos (*pixels*) em uma reta. Em seguida, calcula-se o grau de inclinação formada pela linha identificada. Foram consideradas imagens rotacionadas, aquelas que apresentaram graus de inclinação fora do intervalo [-4, 4] graus.

Tabela 1. Ocorrência das características de interesse na base de dados

Característica	Ocorrência na base	Ocorrência na base (%)
Linhas de texto inclinadas	23 (15)	14,38% (9,38%)
Linhas de texto curvilíneas	30 (10)	18,75% (6,25%)
Variação de iluminação	65 (42)	40,63% (26,25%)
Caracteres pouco nítidos	28 (16)	17,5% (10%)
Ausência das características de interesse	48	30%

A Tabela 1 apresenta a ocorrência de cada característica na base de dados. Observa-se que o maior percentual (40,63%) refere-se a imagens com variação de iluminação. Este percentual decorre pelo fato de que, ao contrário dos *scanners*, que possuem mecanismos especialmente projetados para iluminar a superfície da imagem de maneira uniforme, as fotografias obtidas através de câmeras estão suscetíveis a variações de iluminação do ambiente. A menor incidência refere-se às imagens com linhas de texto inclinadas, são 23 imagens correspondendo a 14,38% das amostras. As imagens que não apresentam nenhuma das características de interesse, representam 30% da base.

É preciso ressaltar que há imagens que apresentam mais de uma característica, visto que a base de dados é constituída de imagens reais. Dessa forma, para realizar o experimento, considerou-se apenas as imagens que possuíam uma única característica de interesse, as quais estão representadas entre parênteses na segunda coluna da Tabela 1.

[1] Algoritmo de autoria de Vecsei (2016). Disponível em: <https://github.com/gaborvecsei/Straighten-Image>

4.1. Configuração e extração

O extrator utilizado no experimento foi o *Tesseract.* A versão utilizada no experimento foi a versão 3.04, compilada e construída fazendo uso do conjunto de ferramentas *Android NDK*[2]. O modo de extração foi definido como *OEM_DEFAULT*, utilizando os arquivos padrão de dados do idioma Português Brasil[3].

Optou-se por utilizar o *Tesseract* em um *smartphone*, objetivando desenvolver uma aplicação cliente capaz de realizar a extração sem a necessidade de requisitar o serviço de extração a um servidor Web.

Após construída e configurada a aplicação, as imagens foram submetidas isoladamente ao *Tesserac*t, armazenando o resultado obtido da extração junto à base de dados.

4.2. Análise

Para avaliar a eficácia da extração, foram adotadas as métricas tradicionais de recuperação de informação: precisão *(precision)*, revocação *(recall)* e F1 *(F-measure).* Para este trabalho, contextualizando a definição de Precisão e Revocação apresentada em [Baeza-Yates and Ribeiro-Neto 2013], assumimos que a precisão mede a fração dos termos recuperados que é relevante e a revocação mensura a fração dos termos relevantes que foi recuperada. Portanto:

$$precisão = \frac{|\,Termos\ relevantes \cap Termos\ recuperados\,|}{|\,Termos\ recuperados|}$$

$$revocação = \frac{|\,Termos\ relevantes \cap Termos\ recuperados\,|}{|\,Termos\ \ relevantes\,|}$$

Um termo relevante é aquele que foi extraído de uma imagem corretamente e, portanto, está presente no gabarito. Por fim, o *F1* consiste na média harmônica entre os índices de precisão e revocação, com o objetivo de fornecer um só índice de medida.

$$F1 = 2 \cdot \frac{precisão \cdot revocação}{precisão + revocação}$$

Ressalta-se que acentuações, distinção de letras maiúsculas e minúsculas, bem como, pontuações foram desconsideradas na análise. Os resultados obtidos são discutidos na seção que segue.

5. Experimento e Resultados

Esta seção apresenta o experimento realizado, bem como os resultados obtidos a partir deste experimento. O objetivo do experimento foi responder a questão: **qual o impacto de cada característica na eficácia da extração?** Para isso foi analisada a F-measure da extração em imagens com as características de interesse.

---

[2] O conjunto de ferramentas Android NDK permite a implementação de partes de aplicativos fazendo uso de linguagens de código nativas como C/C++.

[3] Pacote de treinamento padrão para o idioma Português Brasil para a versão 3.04 do Tesseract. Disponível em: < https://github.com/tesseract-ocr/langdata/tree/master/por >. Acessado em 16 de agosto de 2017.

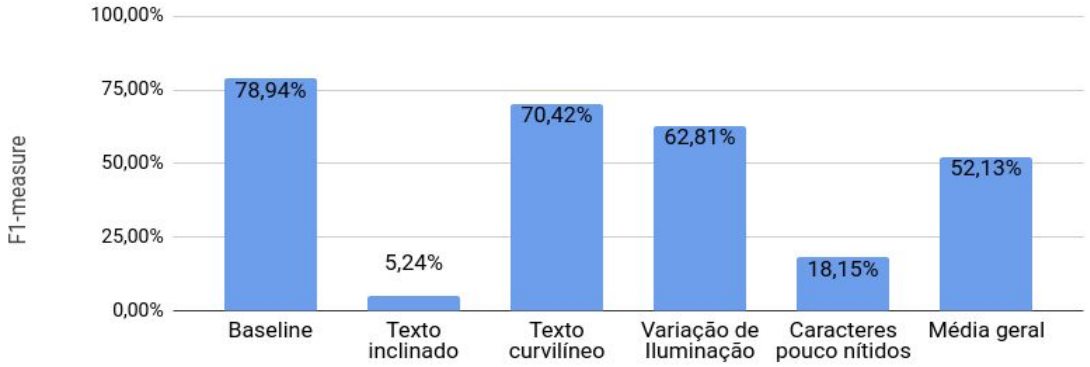

Figura 6: Eficácia da extração em relação às características de interesse

A Figura 6 apresenta os resultados obtidos, onde as colunas representam o percentual de F1. A primeira coluna denominada baseline refere-se às imagens que não possuem nenhuma das características de interesse. A segunda coluna refere-se às imagens com textos inclinados. A terceira coluna diz respeito às imagens que contém textos com aspectos curvilíneos. A quarta coluna apresenta as imagens que possuem variação de iluminação. Imagens que possuem caracteres pouco nítidos estão representados na quinta coluna. Por fim, a sexta coluna apresenta a média geral da eficácia obtida por todas as imagens da base de dados.

Observa-se que as imagens que não possuem nenhuma das características de interesse obtiveram uma média de F1 de 78,94%. As imagens que possuem texto inclinado obtiveram o menor percentual de eficácia (5,24%). Isso se deve ao fato do *Tesseract* não dispor de etapas de pré-processamento para corrigir adversidades relacionadas a inclinação das linhas de texto. Imagens com texto curvilíneo (70,42%) e imagens com variação de iluminação (62,81%), são as características que menos comprometeram a eficácia da extração. A média geral da eficácia da extração alcançou 52,13%, sugerindo que outras características não consideradas neste trabalho podem estar comprometendo a eficácia de extração.

6. Considerações finais

Este trabalho apresentou uma análise da relação entre a eficácia de extração de elementos textuais em imagens e suas características.

Após a realização de experimentos, constatou-se que a inclinação de textos (5,24%), bem como caracteres pouco nítidos (18,15%), são as principais características relacionadas aos baixos índices de eficácia na base de dados utilizada, ao instante que imagens com texto curvilíneo (70,42%) e variação de iluminação (62,81%) não comprometem substancialmente a eficácia de extração.

Como trabalhos futuros, destacam-se: (i) realizar testes em uma base de dados maior; (ii) analisar a influência de outras características; (iii) verificar a relação da eficácia da extração por intervalos de graus de inclinação.

Referências

E. Manica; C. F. Dorneles; R. Galante. (2017). R-Extractor: a method for data extraction from template-based entity-pages. In *Computer Software and Applications Conference (COMPSAC), IEEE 41st Annual*. IEEE. p. 778-787.

A. Labrinidis, H. V. Jagadish. (2012). Challenges and opportunities with big data, *Proceedings of VLDB Endowment*, v. 5, n.12, pp. 2032-2033.

D. Agrawal, P. Bernstein, E. Bertino, et. al. (2012). *Challenges and Opportunities with Big Data - A community white paper developed by leading researchers across the United States*.

Statistic Brain. (2017). Instagram Company Statistics. Disponível em: brain https://www.statisticbrain.com/instagram-company-statistics. Acessado em: 15 de Janeiro de 2018.

D. Berchmans; S. S. Kumar. (2014). Optical character recognition: An overview and an insight. In *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, Kanyakumari, pp. 1361-1365.

N. Islam; Z. Islam; N. Noor. (2016). A Survey on Optical Character Recognition System. Journal of Information & Communication Technology-JICT Vol. 10 Issue.2.

F. Asad et al. (2016) High Performance OCR for Camera-Captured Blurred Documents with LSTM Networks. In *Document Analysis Systems (DAS)*, 2016 12th IAPR Workshop on. IEEE. p. 7-12.

J. Liang, D. Doermann, and H. Li. (2005). Camera-based analysis of text and documents: a survey, International Journal on Document Analysis and Recognition (IJDAR), v. 7, n. 2-3, pp. 84–104.

C. Olah. (2015). Understanding LSTM. Disponível em: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. Acesso em: novembro de 2017.

R. W. Smith. (2017). The Extraction and Recognition of Text from Multimedia Document Images, PhD Thesis, University of Bristol, November 1987.

Tesseract (2015). Tesseract. Disponível em: https://github.com/tesseract-ocr/tesseract. Acesso em: novembro de 2017.

D. M. Kuhn; C. R. Cervi; E. Manica. (2017). Avaliação da eficácia da extração de texto em imagens. VI MOEPEX, IFRS, Campus Ibirubá. Disponível em: https://eventos.ifrs.edu.br/index.php/MoEPExIbiruba/6MOEPEX/paper/view/3332.

OpenCv. (2017). Hough Line Transform. Disponível em: https://docs.opencv.org/2.4/doc/tutorials/imgproc/imgtrans/hough_lines/hough_lines.html. Acessado em: dezembro de 2017.

R. Baeza-Yates; B. Ribeiro Neto. (2013). Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca. Porto Alegre: Bookman Editora.