BID Dataset: a challenge dataset for document processing tasks

Álysson de Sá Soares [1], Ricardo Batista das Neves Junior [1], Byron Leite Dantas Bezerra [1]
[1]Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil
Emails: {alss, rbnj}@ecomp.poli.br, byron.leite@upe.br

***Abstract*—The digital relationship between companies and customers happens through online systems where consumers must upload their identification documents pictures to prove their identities. The existence of this large volume of document images encourages the research development to generate image processing systems to automate tasks usually performed by humans, such as Document Type Classification and Document Reading. The lack of identification documents public datasets delays the research development in document image processing because researchers need to attempt partnerships with private or governmental institutions to obtain the data or build their dataset. In this context, this work presents as main contributions a system to support the automatic creation of identification document public datasets and the Brazilian Identity Document Dataset (BID Dataset): the first Brazilian identification documents public dataset. To accomplish the current personal data privacy law, all information in the BID Dataset comes from fake data. This work aims to increase the velocity of research development in identification document image processing, considering that researchers will be able to use the BID Dataset to develop their research freely.**

***Index Terms*—Identity Document Dataset, Document Segmentation, Document Analysis, Document Recognition.**

I. Introduction

The growth of digital media has led financial and government institutions to engage with their customers through smartphones and other mobile devices [1]. This relationship assumes that the client sends his personal information in text format, and uploads photos of his identification documents to answer the basic and mandatory question of the regulation based on Know Your Customer (KYC): are you who you say you are? [2].

Once the organizations have the images of identification documents of their customers, they can execute some algorithms for the automation of the text field extraction tasks [3], document classification [4], signature extraction [5], in addition to other properties and patterns present in the identification documents images.

Among the approaches for automatic document image processing, algorithms based on Convolutional Neural Networks (CNN) have been widely successfully applied. The first CNN approach to document recognition was proposed by LeCun et al. [6]. Later, Kang et al. [7] presented a CNN for document classification. Tensmeyer [8] presented a CNN approach to binarizing document images.

Personal identification documents contain confidential information that cannot be made public; consequently, there are no datasets of images of identification documents freely available on the internet. In this context, image processing research teams that would like to apply their algorithms to identification documents, need industrial partners who provide the dataset [4], or invest much time and effort to build their own datasets [3] [9]. In particular, due to the Brazilian personal data privacy law [10], there are several restrictions to access, store, and share individual documents.

This lack of availability of identification documents datasets is a severe problem that directly harms this field of research, given that researchers who do not yet have the dataset a priori need more time and effort to build it, in addition to developing the technique applied to the document image processing [11].

In this context, this work presents as main contributions:

1) Public Dataset Identification Documents Generator: an algorithm for generating identification documents datasets. The proposed algorithm can be useful for researchers from around the world to more quickly generate their identification documents datasets.
2) Brazilian Identity Document Dataset (BID Dataset): The first public dataset of Brazilian identification documents. The BID Dataset addresses the problems of Automatic Text Extraction, Optical Character Recognition (OCR) and Document Image Classification. It is worthy to mention all personal data in this dataset are fake.

The next sessions of this work are organized as follows: Session II, presents the previous related works. Session III presents the proposed algorithm for generating datasets from identification documents. Session IV presents the proposed dataset. Session V presents our conclusions and future works.

II. Related Works

Researchers presented the Smartdoc Dataset [12] through the International Conference on Document Analysis and Recognition (ICDAR) competition. The dataset addresses the challenges of OCR and mobile document capture. The dataset contains six types of documents (datasheet, letter, magazine, paper, patent, and tax) from public databases with five documents per class.

The Mobile Identity Document Video dataset (MIDV-500) was proposed by [11] to analyzing and recognizing identity documents on mobile devices. The dataset consists of 500 video clips of 50 different types of identification documents, including 17 types of ID cards, 14 types of passports, 13 types of driving licenses, 6 types of identification documents from

various countries. The dataset is evaluated for facial detection, OCR, and document region detection.

In the ICDAR 2017 competition, researchers presented the DIBCO 2017 [13] dataset. The dataset is composed of images of historical documents in which the dataset's application is focused on the binarization of images of handwritten and printed documents. As a benchmarking proposal, research teams submitted methods for binarizing historical documents.

Dizaj, Shima et al. [14] presented in 2020 a dataset for document corners localization. The dataset contains 1111 images captured in various conditions from smartphones. The images are classified into three categories: (*i*) simple, (*ii*) medium and (*iii*) difficult, in terms of how simple is to identify the position of the document in the image by the eye. The dataset has images with more than 500 different documents of size, material, color, and content, over 500 different backgrounds, in terms of material, texture and design, and have geometric failures.

The processing of images of identification documents has received much attention in the literature. Researchers have presented approaches for identification documents classification [4], automatic handwritten signature segmentation [5], document boundary detection and document text detection [3].

III. Generating a Dataset of Identification Documents with Fake Data

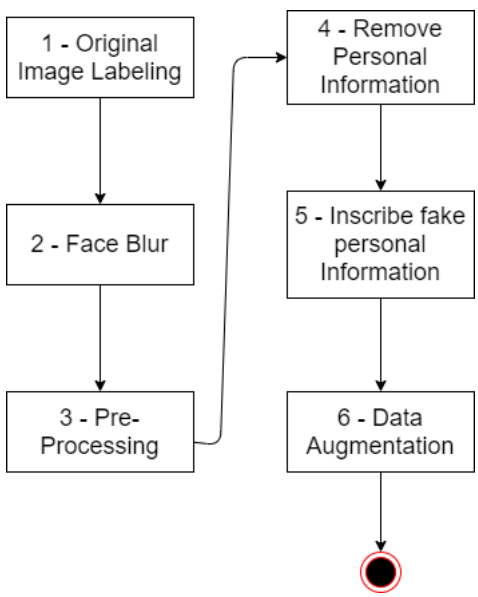

Fig. 1. Flowchart of the proposed algorithm operation

As shown in Figure 1, the proposed algorithm is divided into six main steps, which are detailed below:

1) The original documents images were labeled with the aid of VGG Image Annotator (VIA) [15], an online tool that allows humans to annotate and describe spatial regions in images or video frames and temporal segments in audio or video. With the aid of VIA, we manually annotated the regions where the texts were present in the images. After manual labeling through VIA, we generated a text file with the corresponding texts coordinates, the type of personal information (e.g person's name, date of birth, document number, among others) in those coordinates and the transcription of the original document text in case of non personal information. The original images, together with the text file, were used as the system input.
2) To identify the face region in the document image, we used the Max-Margin Object Detection pre-trained model described in [16] after a blur filter. Once the face region was identified, we hid the individual face through the Blur Gaussian filter [17].
3) In the pre-processing step, all original images were copied to a version with 500 Dots Per Inch (DPI) and double of their sizes. Later, we used the Tesseract-OCR [18] to identify the current orientation of modified images, to keep all original images to the 0º position. This helped the insertion of false personal information horizontally, from left to right (according to the reading direction).
4) Using the coordinates information obtained through manual labeling performed in the first step, we removed personal information by scanning labeled region pixels and replaced the pixel values by Region Dominant Color (DC) (calculated using the Median Cut Algorithm [19]) or by the Average Color (AC) (defined by Equation 1) of three pixels above. If $AC > DC$, AC is applied; otherwise, it is applied DC. As shown in Figure 2, after this process, the resulting image contains only the background and texts of the document layout.

$$AC = \frac{px_x^{y-1} + px_x^{y-2} + px_x^{y-3}}{3} \tag{1}$$

where px represents the pixel in processing, y is the y-axis coordinate and x is the x-axis coordinate.
5) We generated fake information for all fields present in the document (as name, date of birth, affiliation, document number, among others). The fake names to be inscribed in the documents were randomly selected from two lists. The first list is a text file composed of fake names, and the second list is composed of fake Brazilian surnames. As a source for generating the fake information of cities and provinces, we used the data from Instituto Brasileiro de Geografia e Estatística (IBGE) [20] and randomly got a province, then we took a city. To produce the numeric data fields (as the document number, date of bird, among others), we used algorithms for randomly generate numbers with the field size. From these fake data, we made a mask with the same dimensions of the original image, where the fake text is represented by black pixels, the background is represented by white pixels, as shown in Figure 3. The fake personal information was stored in a .txt file to serve as Ground Truth (GT) for Optical Character Recognition (OCR) methods.

6) Finally, in the Data Augmentation stage, we performed rotation operations, modifying brightness and contrast, applying blur, and random noise, among others, to build new images for the base.

The source code of the proposed method previously described is available on the following link: https://github.com/AlyssonDSS/GeradorBaseSintetica.

Fig. 2. Document Image after removing personal information.

Fig. 3. Mask with false personal information.

IV. Brazilian Identity Document Dataset (BID Dataset)

Brazilian Identity Document Dataset (BID Dataset) aims at three crucial challenges in the Computer Vision field: (*i*) Document Images Classification; (*ii*) Text Region Segmentation and (*iii*) Optical Character Recognition (OCR). BID Dataset is composed of Brazilian ID documents divided into eight classes: front and back faces of National Driver's License (CNH), CNH front face, CNH back face, Natural Persons Register (CPF) front face, CPF back face, General Registration (RG) front face, RG back face, and RG front and back faces.

As shown in Figure 4, the dataset consists of an original image (document photo), and two Ground Truth (GT) types.

(a) Original Image (b) GT for Text Segmentation challenge

```
x, y, width, height, transcription
157, 124, 48, 14, NOME
146, 143, 206, 10, ELKI PRECIOSO MANFRED
458, 180, 241, 12, DOC. IDENTIDADE/ÓRG EMISSOR UF
451, 197, 86, 11, 683139378
645, 200, 39, 10, SPTC
772, 200, 24, 11, MG
458, 236, 30, 13, CPF
444, 252, 151, 12, 279.201.222-60
681, 235, 124, 14, DATA NASCIMENTO
667, 253, 127, 16, 07/09/2007
462, 287, 63, 16, FILIAÇÃO
446, 309, 251, 11, MAGDALENA BOTOSSI LORET
446, 330, 107, 12, CAVALCANTI
464, 377, 48, 10, SILVA
453, 430, 74, 18, PERMISSÃO
649, 433, 34, 14, ACC
745, 432, 57, 17, CAT. HAB.
758, 454, 12, 13, C
78, 368, 23, 183, 024460530
41, 321, 22, 222, O TERRITÓRIO NACIONAL
19, 352, 21, 153, VÁLIDA EM TODO
147, 493, 88, 14, N° REGISTRO
178, 510, 134, 13, 36089010795
428, 494, 67, 15, VALIDADE
429, 513, 116, 16, 21/09/1987
617, 494, 104, 16, 1ª HABILITAÇÃO
629, 514, 127, 17, 07/04/2017
150, 570, 98, 16, OBSERVAÇÕES
307, 866, 172, 15, ASSINATURA DO PORTADOR
153, 888, 46, 16, LOCAL
137, 911, 128, 13, PIO XII-MA
17, 888, 22, 210, PROIBIDO PLASTIFICAR
66, 904, 28, 209, 644126533
635, 890, 95, 17, DATA EMISSÃO
647, 914, 105, 13, 25/06/1951
316, 951, 143, 12, DAENEKAS RIE
311, 1013, 160, 11, ASSINATURA DO EMISSOR
643, 964, 117, 13, 30066801050
639, 988, 107, 12, PI300032470
[253, 253, 726, 727], [1040, 1072, 1075, 1039], -1, -1, DETRAN - MG (MINAS GERAIS)
143, 611, 23, 16, J;
165, 25, 657, 27, REPÚBLICA FEDERATIVA DO BRASIL
308, 54, 368, 19, MINISTÉRIO DAS CIDADES
205, 75, 574, 20, DEPARTAMENTO NACIONAL DE TRÂNSITO
209, 99, 564, 24, CARTEIRA NACIONAL DE HABILITAÇÃO
449, 354, 271, 12, PARONI GIANEZ SKALINSKI
```

(c) GT for OCR challenge

Fig. 4. BID Dataset: Image Sample

To address the segmentation challenge, the first GT type available is the image mask (in black pixels), delimiting the text regions (in white pixels), as shown in Figure 4 (b). To address the OCR challenge, the second GT type is the image content transcription, and the text regions coordinates, as shown in Figure 4 (c).

BID Dataset is composed of 28,800 document images, with 3,600 samples for each class. The dataset presents a variety of images in different resolutions, lighting conditions, perspectives, orientations, contrasts, and noise. The personal

information in the document is false and was generated according to section III.

The proposed dataset can be downloaded from the following repository: https://github.com/ricardobnjunior/Brazilian-Identity-Document-Dataset.

V. Conclusion and Future Works

Digital relationship between companies and customers has boosted personal online registration in corporate systems, in which customers need to prove their identifications by uploading their documents photos. This fact encourages the researches development in image processing of identification documents to support the automation of different tasks such as Automatic Text Segmentation, Document Type Classification, Optical Character Recognition, among others.

Given the Brazilian personal data privacy law [10] and similar regulations in other countries in the World, there are several restrictions to access, store, and share individual documents. Therefore, the existing datasets are private, limiting advances in the document processing field, since the researchers do not have access to these datasets to evaluate and improve their algorithms.

In this context, the main contributions of this work are: (i) a powerful algorithm to generate public identification documents datasets with non-confidential and private data; and (ii) the first open Brazilian public dataset of identification documents, named Brazilian Identity Document Dataset (BID Dataset), without personal and sensitive data, in attention to the Brazilian personal data privacy law [10]. The proposed algorithm is divided into six main stages: Original Image Labeling; Face Blur; Blur Application on People's Faces; Preprocessing; Personal Information Removal; Insertion of False Personal Information and Data Augmentation. The presented dataset is composed of 28,000 images, divided into eight classes.

Both contributions of our work will be freely available and must be used only for research purposes. Therefore, we hope the proposed system can support researchers around the world to generate their identification document datasets to develop their research without the need for partnerships with big private or governmental institutions. While the BID Dataset is useful for researchers to try and improve their document image processing algorithms taking into account the challenges of this brand new dataset.

As future work, the authors will investigate image processing algorithms based on Deep Learning to address the possible challenges proposed by the BID Dataset.

Acknowledgment

This study was financed in part by: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Fundação de Amparo a Ciência e Tecnologia do Estado de Pernambuco (FACEPE), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - Brazilian research agencies.

References

[1] K. Gai, M. Qiu, and X. Sun, "A survey on fintech," *Journal of Network and Computer Applications*, vol. 103, pp. 262–273, 2018.

[2] R. R. Mullins, M. Ahearne, S. K. Lam, Z. R. Hall, and J. P. Boichuk, "Know your customer: How salesperson perceptions of customer relationship quality form and influence account profitability," *Journal of Marketing*, vol. 78, no. 6, pp. 38–58, 2014.

[3] R. B. das Neves Junior, L. F. Verçosa, D. Macêdo, B. L. D. Bezerra, and C. Zanchettin, "A fast fully octave convolutional neural network for document image segmentation," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[4] R. Sicre, A. M. Awal, and T. Furon, "Identity documents classification as an image classification problem," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 602–613.

[5] C. A. Lopes Junior, M. H. M. da Silva, B. L. D. Bezerra, B. J. T. Fernandes, and D. Impedovo, "Fcn+ rl: A fully convolutional network followed by refinement layers to offline handwritten signature segmentation," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[7] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for document image classification," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 3168–3172.

[8] C. Tensmeyer and T. Martinez, "Document image binarization with fully convolutional neural networks," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 99–104.

[9] A. M. Awal, N. Ghanmi, R. Sicre, and T. Furon, "Complex document classification and localization application on identity document images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 426–431.

[10] P. da República do Brasil, "Lei geral de proteção de dados pessoais (lgpd)," http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm, 2018, [Online; accessed 2020-07-17].

[11] V. V. Arlazarov, K. B. Bulatov, T. S. Chernov, and V. L. Arlazarov, "Midv-500: a dataset for identity document analysis and recognition on mobile devices in video stream," , vol. 43, no. 5, 2019.

[12] J.-C. Burie, J. Chazalon, M. Coustaty, S. Eskenazi, M. M. Luqman, M. Mehri, N. Nayef, J.-M. Ogier, S. Prum, and M. Rusiñol, "Icdar2015 competition on smartphone document capture and ocr (smartdoc)," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 1161–1165.

[13] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "Icdar2017 competition on document image binarization (dibco 2017)," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 1395–1403.

[14] S. B. Dizaj, M. Soheili, and A. Mansouri, "A new image dataset for document corner localization," in *2020 International Conference on Machine Vision and Image Processing (MVIP)*. IEEE, 2020, pp. 1–4.

[15] A. Dutta and A. Zisserman, "The via annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2276–2279.

[16] D. E. King, "Max-margin object detection," *arXiv preprint arXiv:1502.00046*, 2015.

[17] E. S. Gedraite and M. Hadad, "Investigation on the effect of a gaussian blur in image filtering and segmentation," in *Proceedings ELMAR-2011*. IEEE, 2011, pp. 393–396.

[18] R. Smith, "An overview of the tesseract ocr engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, vol. 2. IEEE, 2007, pp. 629–633.

[19] P. Heckbert, "Color image quantization for frame buffer display," *ACM Siggraph Computer Graphics*, vol. 16, no. 3, pp. 297–307, 1982.

[20] I. B. de Geografia e Estatística, "Censo demográfico — ibge," https://www.ibge.gov.br/estatisticas/sociais/populacao/22827-censo-2020-censo4.html?=&t=downloads, 2020, [Online; accessed 2020-07-07].