13th International Conference on Computer and Knowledge Engineering (ICCKE 2023), November 1-2, 2023, Ferdowsi University of Mashhad, Iran

SUT: a new multi-purpose synthetic dataset for Farsi document image analysis

2023 13th International Conference on Computer and Knowledge Engineering (ICCKE) | 979-8-3503-3015-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCKE60553.2023.10326243

Elham Shabaninia
Department of Applied Mathematics,
Faculty of Sciences and Modern Technologies,
Graduate University of Advanced Technology,
Kerman, Iran
e.shabaninia@kgut.ac.ir

Fatemeh sadat Eslami
Department of Computer Engineering,
Sirjan University of Technology,
Sirjan, Iran
fatemesslm@gmail.com

Ali Afkari-Fahandari
Department of Electrical Engineering,
Shahid Bahonar University of Kerman,
Kerman, Iran
aliafkari74@eng.uk.ac.ir

Hossein Nezamabadi-pour
Department of Electrical Engineering,
Shahid Bahonar University of Kerman,
Kerman, Iran
nezam@uk.ac.ir

***Abstract*— This paper introduces a new large-scale dataset for Farsi document images, named SUT, which aims to tackle the challenges associated with obtaining diverse and substantial ground-truth data for supervised models in document image analysis (DIA) tasks, like document image classification, text detection and recognition, and information retrieval. The dataset comprises 62,453 images that have been categorized into 21 distinct classes, including identity documents featuring synthetically generated personal information superimposed on various backgrounds. The dataset also includes corresponding files with labeling information for the images. The ground-truth data is organized in CSV files containing image file paths and associated information about the embedded data. To demonstrate the efficacy of the SUT dataset in DIA tasks, it was utilized for document classification (achieving an accuracy of 86% using a convolutional neural network) and OCR (achieving a CER of 0.083 and 0.072 using Tesseract and EasyOCR engines, respectively). The SUT dataset serves as an esteemed asset for scholars engaged in the development and assessment of supervised models in Farsi document image analysis.**

Keywords—Farsi database, Document Image Analysis, Document Classification, Optical Character Recognition.

I. INTRODUCTION

As the volume of digitized documents continues to grow, there is a rising requisition for automated and intelligent systems capable of understanding the logical structure of documents. This is essential for the efficient management of these documents. Document image analysis (DIA) involves techniques used to transform pixel data from images of documents into a computer-readable format [1]. In intelligent document management systems, document image analysis is essential for efficient document archiving, retrieval, information mining, and so on. Document image analysis tasks usually involve different steps.

Layout analysis and automatic classification of document images typically serve as the initial steps for DIA. Document layout analysis aims to identify the relevant elements in a document image, which may include groups of text-lines, figures, tables, mathematical symbols, etc. [2, 3]. Nevertheless, the goal of document image classification is to classify a given document image into pre-defined categories. The performance of a system can be elevated through the accurate initial categorization of input documents into predetermined classes. Layout analysis and image classification results may complement each other. Knowing the category to which the document belongs can aid in recognizing the components of the document image. On the other hand, documents can be categorized based on their textual content, structural information, and visual structure [4, 5]. Document classifiers that rely on textual content typically utilize optical character recognition (OCR) for text extraction from document images, followed by natural language processing (NLP) approaches to classify the text. Although there are recent advances in text classification using NLP, the OCR is still a challenging task [6], particularly for languages like Farsi [7]. This is due to the complexities of the Farsi language, including connected characters, significant similarity between different characters, dots, semicircles, sub-words, etc. Hence, the effectiveness of such systems depends on the successful optical character recognition (OCR) and subsequent accurate classification of text using NLP. The classification task becomes more complex when relying on structural information is also challenging, as the documents are not always well-structured and the layout may lack distinctiveness. In fact, certain classes may exhibit highly similar structures.

Recently, document image analysis (DIA) using deep learning approaches has attracted much attention [8, 9]. In fact, deep CNN techniques have become widely adopted for vast image

classification datasets, like ImageNet [10], achieving superior results that considerably surpass traditional methods. Recently, the transformer-based models [11] have shown that a pure transformer model can achieve superior results on different image and video recognition tasks, without requiring convolutional blocks [12].

This paper presents a dataset of Farsi document images that includes 21 distinct classes of commonly used documents in office automation, distributed enrollments, etc. The dataset includes identity-related document classes, namely "national card," "birth certificate," "passport," etc., which are particularly challenging to obtain in large quantities due to privacy restrictions. The information about persons is synthetically superimposed on images. Moreover, the accompanying CSV files specify the embedded information and their location within the images. To introduce variability in the data, the dataset incorporates different backgrounds that are commonly encountered while working with documents. In summary, the dataset's contributions can be outlined as follows:

1- The SUT dataset is the first comprehensive dataset developed explicitly for Farsi document analysis, representing a significant contribution to the field[1].
2- The SUT dataset is versatile and can be employed for numerous tasks, including document image classification, text detection, OCR, information retrieval, and more.
3- The SUT dataset includes a diverse range of identity-related document classes, such as national cards, birth certificates, and passports, making it a beneficial asset for practitioners and researchers working with these types of documents.
4- The performance of the SUT dataset has been evaluated on document classification and OCR tasks, demonstrating its effectiveness as a multi-purpose dataset.

II. RELATED WORKS

To enable fair comparisons among different methods in document image classification, public datasets are highly desirable, yet there are only a handful number of such datasets currently available. IIT CDIP Test Collection [13] dataset consists of high-resolution images from more than seven million scanned documents, which were gathered from public records related to legal cases against American tobacco companies. The documents are manually labeled with tags, and in many cases, the first tag of a document image denotes its category. However, a significant number of documents in the dataset have tags that are either missing or incorrect [8]. Tobacco-3482 (SmallTobacco) dataset [14] comprising of 3482 images from the IIT collection, each labeled with one of ten categories. The distribution of images across categories is uneven, with the "letter" category possessing the highest number of images. The dataset is partitioned into training, validation, and testing sets as described in [14], with 800 images used for training, 200 for validation, and the rest for testing. Since the dataset is relatively small, 10 random splits in those proportions were created and results represent the median performance across those splits. The RVL-CDIP dataset, also known as BigTobacco [8], involves 400,000 annotated images from 16 categories in the IIT CDIP collection, with 25,000 images randomly sampled from each category. The dataset is divided into training, validation, and testing sets with ratios similar to those of ImageNet, where 320,000, 40,000, and 40,000 images were used for training, validation, and testing, respectively. RVL-CDIP can be regarded as the counterpart of ImageNet within the document image community. It should be noted that the dataset contains a significant amount of noise and variance in the composition of each document class. NIST Special Database 2 (NIST-SPDB2) [15] dataset contains 5,590 machine-printed tax form images. The images are binary and were synthesized to replicate the appearance of scanned user-filled forms. The dataset is divided into 20 classes based on the types of forms. NIST Special Database 6 (NIST-SPDB6) [16] consists of 5,595 synthesized binary tax form images. The dataset is divided into 20 classes, similar to NIST-SPDB2. However, unlike NIST-SPDB2, the tax forms in NIST-SPDB6 are handwritten rather than machine-printed. Medical Article Records Ground truth (MARG) [17] dataset comprises of 1,553 document images, which were collected from the first pages of medical journals. These images are partitioned into 9 different classes. Although identity documents hold valuable information, publicly available datasets for them are generally hard to come by due to regulatory restrictions around the world. However, there are a few datasets available, one of which is the Brazilian Identity Document (BID) Dataset [18], which contains 28,800 synthetic images of eight different document types. The dataset was created by overlaying text field values on masked document regions, with the faces of the document owners blurred to protect their privacy. Another useful dataset is the Mobile Identity Document Video dataset family (MIDV) [19]. The first dataset in this family, MIDV-500 [19], comprises 500 video clips featuring 50 identity documents captured under five different settings, including hand, keyboard, table, clutter, and partial. MIDV-2019 [20] extends the MIDV-500 by capturing video clips under very low lighting environments and higher projective distortions. The most recent addition, MIDV-2020 [21], was built upon the 10 document types included in MIDV-500 and MIDV-2019, with the aim of providing diversity in text fields, signatures, and faces while maintaining the dataset's realistic nature.

Compared to the existing datasets, the SUT dataset is a comprehensive dataset specifically designed for Farsi document analysis. It contains various identity-related document categories, including national cards, birth certificates, passports, etc. In addition, the dataset has a wide range of potential uses and can be applied to various tasks, including document image classification, text detection, OCR, and information retrieval.

[1] https://github.com/aliiafkari/SUT_Dataset

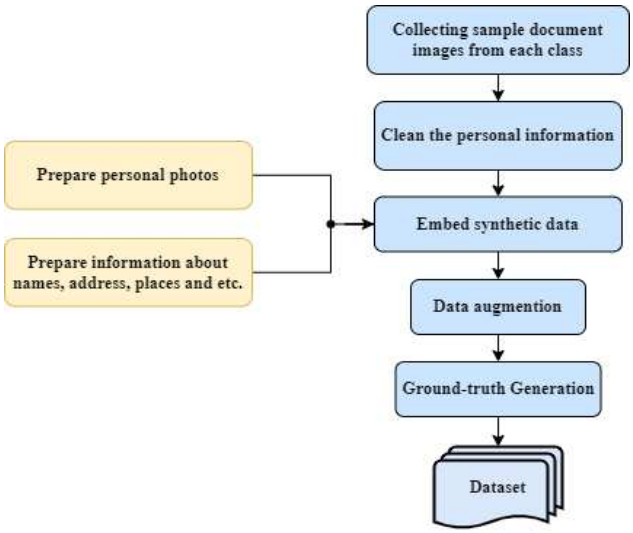


Fig. 1. The process of dataset generation.

III. DATASET GENERATION

An overview of the process used to create the dataset is depicted in Fig. 1, which involves four main steps. The first step is collecting sample document images. The second step involves removing any personal information from the documents to ensure data privacy. In the third step, personal data is embedded into the documents for identification purposes. Finally, in the fourth step, data augmentation techniques are applied to enhance the diversity of available data for training and testing the system.

A. Collecting Sample Document Images

As mentioned before, the SUT dataset contains 21 classes of Farsi document images. Firstly, it is needed to gather a limited number of images for each class. These images are then cleaned and used as background for embedding data if required. Note that for some classes such as "paper", "newspapers", and "magazines" no data is embedded.

B. Cleaning the Personal information

Identity documents, such as identity cards or driver's licenses, often contain sensitive personal information, including a person's national identification number, date of birth, and full name. To mitigate the risk of unauthorized access or misuse of this information, it is essential to remove all personal data from the images obtained during the document collection process. This process, known as data cleaning, ensures that the images are free from sensitive information and ready for further processing. Once personal data has been removed, the images can be safely used for embedding personal data in the next step. Fig. 2 illustrates the result of the data-cleaning process for identity card category.

Fig. 2. Identity card after cleaning personal data.

C. Embedding Synthetic Personal Data

This step aims to embed synthetic data across various categories. In some document categories, personal information, such as a photograph of the document owner, needs to be embedded within the image. To meet this requirement, a collection of 50 personal photos, featuring both men and women, was gathered from the internet. However, in Iran, women are required to cover their hair in accordance with the hijab-wearing guidelines. Therefore, the images of women were edited to include landmarks that include various types of headscarves, as depicted in Fig. 3. In addition to personal photos, documents contain various types of information, such as names, numbers, addresses, and so on. To create a realistic image, all of the data must be consistent with each other. For instance, in certain documents such as passports, names are written in both Farsi and English, and these two strings should match each other. Another example is the consistency between the names and the gender of the individuals depicted in the personal photos.

D. Data Augmentation

In practice, document images may be captured on a variety of surfaces, such as tables, floors, and carpets. Inspired by this, a range of background images are integrated into the images of the documents, as demonstrated in Fig. 4. This process aims to improve the quality of the dataset by providing a more accurate reflection of real-world scenarios.

E. Ground-truth Generation

To obtain the ground-truth data, a CSV file is created with six columns for every category, as depicted in Fig. 5. The name of the image is indicated in the first column, while column two to five provide details on the placement coordinates of the text and subsequently its length and height. Finally, the synthesized text within the image is stored in the last column. This information helps to find the coordinates of embedded data in the images and is vital for tasks such as text detection.

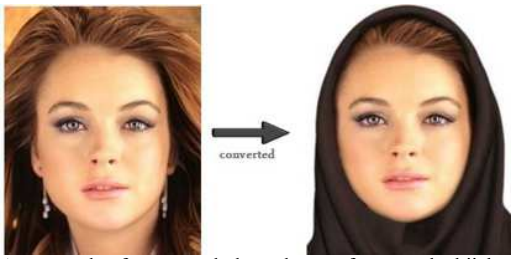


Fig. 3. An example of a personal photo that conforms to the hijab-wearing guideline for women in Iran.

Fig. 4. Sample underlying images used for data augmentation.

	A	B	C	D	E	F
1	filename	x_coord	y_coord	W	H	text
2	men_jadid1	675	216	230	40	6304404347
3	men_jadid1	800	300	105	35	کامشاد
4	men_jadid1	821	370	84	35	امیری
5	men_jadid1	723	438	182	35	1341 / 4 / 11
6	men_jadid1	849	505	56	35	کیا
7	men_jadid1	723	565	182	35	1410 / 2 / 22
8	men_jadid2	675	216	230	40	9675510190
9	men_jadid2	845	300	60	35	کیا
10	men_jadid2	793	370	112	35	انتظامی
11	men_jadid2	723	438	182	35	1360 / 8 / 14
12	men_jadid2	821	505	84	35	فرجاد
13	men_jadid2	709	565	196	35	1408 / 11 / 27
14	men_jadid3	675	216	230	40	3550134011
15	men_jadid3	800	300	105	35	فرخزاد
16	men_jadid3	807	370	98	35	اشراقی
17	men_jadid3	709	438	196	35	1350 / 12 / 29
18	men_jadid3	821	505	84	35	مهرنگ
19	men_jadid3	737	565	168	35	1410 / 2 / 4

Fig. 5. An example of ground-truth data.

Fig. 6 shows sample images generated in the SUT dataset. As this figure shows the images are realistic and resemble the original images. Additionally, Table 1 shows the distribution of generated images in each class with and without background (underlying image). As this table shows more images in this dataset are with background images.

IV. EVALUATION RESULTS

As previously mentioned, the SUT dataset is a multi-purpose dataset that can be utilized for various tasks. In light of this, we have chosen to evaluate the dataset's performance in document classification and OCR tasks in the following subsections.

A. SUT Dataset for Document Classification Task

This section focuses on utilizing the SUT dataset for document image classification. To accomplish this task, a pre-trained convolutional neural network (VGG16) is employed. The VGG16 model is used without the top classification layers, and three fully connected layers are added for fine-tuning. The RMSprop optimizer and categorical cross-entropy loss are employed for training, while the VGG16 backbone weights are kept frozen. The dataset is partitioned into training, validation, and testing portions, with proportions of 70%, 15%, and 15%, respectively. Training the model takes place over 10 epochs, with the batch size specified as 32. Upon evaluation, the model achieves an accuracy of 86% on the testing data. Fig. 7 depicts the loss and accuracy during training and evaluation.

Fig. 6. Sample generated images in each class.

TABLE I. THE NUMBER OF GENERATED IMAGES IN EACH CLASS WITH/WITHOUT BACKGROUND

Class name	Number of images	
	With background	*Without background*
Birth Certificate1_page1	2,400	1,600
Birth Certificate1_page2	1,173	250
Birth Certificate2_page1	3,584	1,200
Birth Certificate2_page2	1,800	600
Birth Certificate2_page3	1,200	300
Birth Certificate2_page4	2,100	800
Identity Card1	1,782	436
Identity Card2	2,160	461
Passport	3,600	1,200
Military Service Card1	1,500	500
Military Service Card2	3,000	504
Bank Document1	1,200	400
Bank Document2	1,200	402
Bank Document3	1,890	400
Bank Document4	3,345	630
Electricity Bill	2,828	354
Water bill	2,784	426
Gas Bill	3,000	1,000
Magazine	2,859	648
Newspaper	2,937	504
Paper	3,000	496
Total	49,342	13,111
	62,453	

Fig. 8 depicts the confusion matrix for the document image classification using the SUT dataset. The matrix indicates that the majority of classes are correctly classified, although certain classes exhibit confusion with each other.

B. SUT Dataset for Optical Character Recognition Task

In this section, the capabilities of the SUT dataset for Farsi printed optical character recognition are investigated. As previously mentioned, the SUT dataset is capable of being utilized for text recognition and detection tasks, using the provided information as ground-truth data. For simplicity, evaluation is limited to two classes from the dataset, specifically "identity card1" and "identity card2," to generate images containing information such as names, last names, birth dates, and national numbers. It is noteworthy that this technique can be applied to other classes as well. The CSV file provides coordinates that can be utilized to crop each image and extract a new image containing a word or string of numbers. These images are especially suitable for evaluating the efficiency of OCR methods. The number of generated images for the OCR task is represented in Table 2.

To assess the effectiveness of the dataset, two commonly used OCR engines, Tesseract [22] and EasyOCR [23] are utilized. The character error rate (CER) metric is employed to evaluate the results, where it quantifies the accuracy of character recognition by comparing the count of correctly identified characters to the total character count. The results indicate a CER of 0.083 and 0.072 for Tesseract and EasyOCR, respectively. Table 3 shows the OCR results for some sample images along with the ground-truth text. It is worth noting that some test images contain noise and background. Despite being trained on limited Farsi fonts, both engines struggle with recognizing strings of numbers. Nevertheless, EasyOCR outperforms Tesseract in terms of overall accuracy.

V. CONCLUSIONS

Document image analysis (DIA) is a critical component in various applications, such as office automation, document archiving, digital libraries, and distributed enrollment systems. However, developing supervised models for DIA tasks requires large quantities of Persian document images that can serve as ground-truth data. This task is particularly challenging when it comes to identity documents due to privacy restrictions that limit the availability of personal data. To overcome these challenges, this paper proposed a novel large-scale multi-purpose dataset for Farsi document images, which we call the SUT dataset.

The SUT dataset is designed to provide a comprehensive and diverse collection of document images that can facilitate research in various DIA tasks, such as document classification, OCR, and text detection. The dataset includes a wide range of document types, such as passports, national ID cards, driving licenses, and birth certificates. To ensure the privacy of individuals, sensitive information has been removed from the documents.

The results obtained from evaluating the SUT dataset in document classification and OCR tasks demonstrate its effectiveness and potential for use in future research. The SUT dataset is an important asset for practitioners and researchers in the area of document analysis, enabling them to develop and evaluate models that can handle a wide range of Farsi documents effectively.

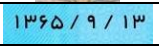
ACKNOWLEDGMENT

The authors wish to acknowledge that the origins of this dataset can be traced back to the computer vision course taught by the first author at Sirjan University of Technology (SUT). They are grateful for the assistance provided by the students who contributed some of the images as part of their assignments in that class.

TABLE II. THE NUMBER OF GENERATED IMAGES ONLY FROM IDENTITY CARDS AS SAMPLES FOR OPTICAL CHARACTER RECOGNITION.

	STRING OF NUMBERS	DATE	NAME
NUMBER OF IMAGES	1,333	1,358	2,689
TOTAL	5,380		

TABLE III. OCR RESULTS USING TESSERACT AND EASYOCR ENGINES.

Ground-truth	Predicted text	
	Tesseract	*EasyOCR*
	برزن	برزن
	یاهتر	باهنر
	13-5 3	۱۳٦٥/۹/۱۳
	330550 1	3555 1 1 5 34

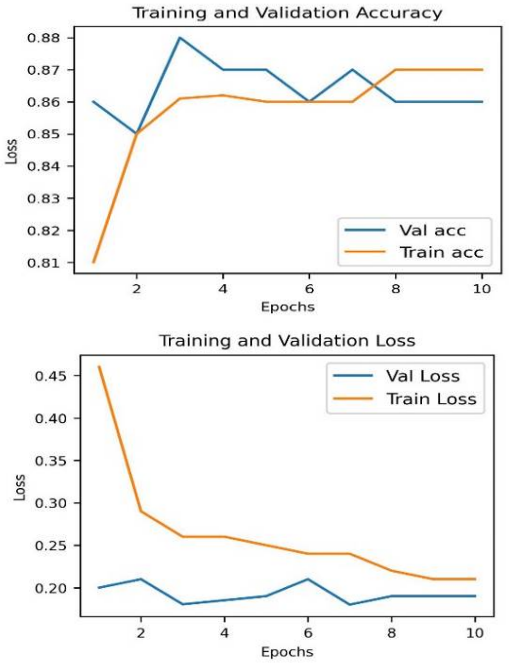


Fig. 7. Accuracy and loss for training and validation of document image classification.

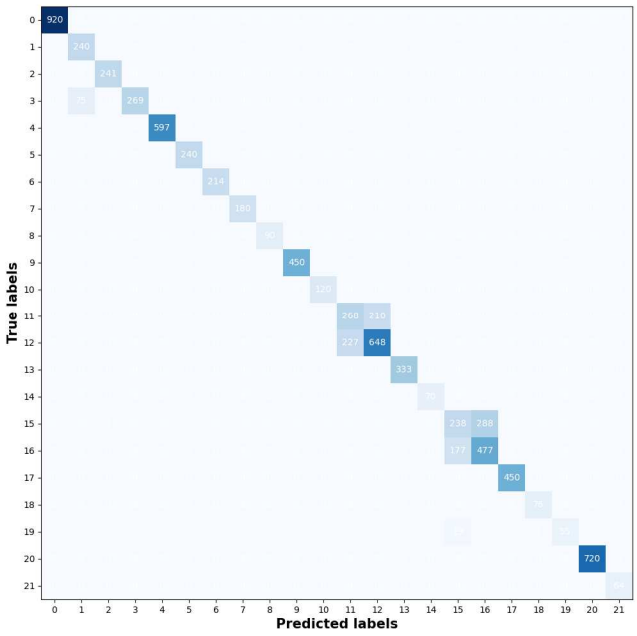


Fig. 8. The confusion matrix for document image classification using the SUT dataset.

REFERENCES

[1] R. Kasturi, L. O'gorman, and V. Govindaraju, "Document image analysis: A primer," *Sadhana,* vol. 27, pp. 3-22, 2002.

[2] L. Pisaneschi, A. Gemelli, and S. Marinai, "Automatic generation of scientific papers for data augmentation in document layout analysis," *Pattern Recognition Letters,* vol. 167, pp. 38-44, 2023.

[3] G. M. Binmakhashen and S. A. Mahmoud, "Document layout analysis: a comprehensive survey," *ACM Computing Surveys (CSUR),* vol. 52, no. 6, pp. 1-36, 2019.

[4] L. Liu, Z. Wang, T. Qiu, Q. Chen, Y. Lu, and C. Y. Suen, "Document image classification: Progress over two decades," *Neurocomputing,* vol. 453, pp. 223-240, 2021.

[5] N. Chen and D. Blostein, "A survey of document image classification: problem statement, classifier architecture and performance evaluation," *International Journal of Document Analysis and Recognition (IJDAR),* vol. 10, pp. 1-16, 2007.

[6] F. sadat Hosseini, S. Kashef, E. Shabaninia, and H. Nezamabadi-pour, "Idpl-pfod: an image dataset of printed Farsi text for OCR research," in *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*, 2021, pp. 22-31.

[7] S. Kashef, H. Nezamabadi-pour, and E. Shabaninia, "A review on deep learning approaches for optical character recognition with emphasis on Persian, Arabic and Urdu scripts," *Journal of Machine Vision and Image Processing,* vol. 8, no. 4, pp. 51-85, 2021.

[8] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015: IEEE, pp. 991-995.

[9] 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*, 2014: IEEE, pp. 3168-3172.

[10] W. Wang, Y. Yang, X. Wang, W. Wang, and J. Li, "Development of convolutional neural network and its application in image classification: a survey," *Optical Engineering,* vol. 58, no. 4, pp. 040901-040901, 2019.

[11] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.

[12] E. Shabaninia, H. Nezamabadi-pour, and F. Shafizadegan, "Transformers in Action Recognition: A Review on Temporal Modeling," *arXiv preprint arXiv:2302.01921,* 2022.

[13] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard, "Building a test collection for complex document information processing," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 665-666.

[14] J. Kumar and D. Doermann, "Unsupervised classification of structurally similar document images," in *2013 12th International Conference on Document Analysis and Recognition*, 2013: IEEE, pp. 1225-1229.

[15] O. A. https://www.nist.gov/srd/nist-special-database-2.

[16] N. S. D. O. A. https://www.nist.gov/srd/nist-special-database-6.

[17] G. Ford and G. R. Thoma, "Ground truth data for document image analysis," in *Symposium on Document Image Understanding and Technology*, 2003, pp. 199-205.

[18] A. de Sá Soares, R. B. das Neves Junior, and B. L. D. Bezerra, "BID Dataset: a challenge dataset for document processing tasks," in *Anais Estendidos do XXXIII Conference on Graphics, Patterns and Images*, 2020: SBC, pp. 143-146.

[19] 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," *Компьютерная оптика,* vol. 43, no. 5, pp. 818-824, 2019.

[20] K. Bulatov, D. Matalov, and V. V. Arlazarov, "MIDV-2019: challenges of the modern mobile-based document OCR," in *Twelfth International Conference on Machine Vision (ICMV 2019)*, 2020, vol. 11433: SPIE, pp. 717-722.

[21] B. K. Bulatovich *et al.*, "MIDV-2020: a comprehensive benchmark dataset for identity document analysis," *Компьютерная оптика,* vol. 46, no. 2, pp. 252-270, 2022.

[22] R. Smith, "An overview of the Tesseract OCR engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, 2007, vol. 2: IEEE, pp. 629-633.

[23] A. Jaided, "EasyOCR," *Retrieved October,* vol. 9, no. 2020, p. 5, 2020