Comparative Evaluation of Text Recognition Methods for Field Extraction in Brazilian National Driver's Licenses

Suziane L. R. R. Mattos, Francisco de A. Boldt, Thiago M. Paixão

[1]Instituto Federal do Espírito Santo (IFES)
Serra – ES – Brazil

suzianeramalho@gmail.com, {franciscoa, thiago.paixao}@ifes.edu.br

Abstract. *With the advancement of digitalization, the demand for document information extraction solutions has grown, but studies focusing on Brazilian Driver's Licenses (CNH) remain scarce. This paper highlights this gap by evaluating OCR methods, including TrOCR (with varying settings), Tesseract, and GPT models, for extracting information from CNH documents. The study considers challenges such as lighting and variable image quality, using metrics such as error rate and recognition evaluation in specific fields. The results demonstrate promising performance using open source tools, offering insights into the advantages and limitations of each model.*

1. Introduction

The increasing use of digital services has been a constant trend in recent decades, with users predominantly interacting through online platforms [Gov 2024]. In this context, institutions such as banks, e-commerce platforms, social networks, and government agencies often request document images to verify users' identities [Castelblanco et al. 2020], granting them access to platform services. For scalable and reliable identification, automating information extraction from identity documents is essential [Baviskar et al. 2021]. Computational tools enable efficient processing of large data volumes, reducing response times and costs associated with manual analysis. To achieve robust recognition, deep learning models – specifically convolutional neural networks (CNNs) and Transformer-based models [Vaswani et al. 2017] – have been employed due to their superior performance in tasks such as text detection and optical character recognition (OCR) [Subramani et al. 2021].

Deep learning-based solutions have been widely proposed for processing identity documents. Attivissimo et al. 2019 used CNNs and a single-line text recognition model to classify and extract information from Italian documents based on synthetic data. Similarly, Chandra and Stefanus 2021 implemented an OCR pipeline for Indonesian documents, combining CNNs to detect and recognize specific fields. Hoai et al. 2021 developed a recognition system for Vietnamese documents using convolutional recurrent neural networks, achieving higher accuracy than commercial solutions when handling complex texts. More recently, Wojcik et al. 2023 introduced the NBID synthetic dataset of official documents, along with a synthesizer method, and evaluated it using three end-to-end visual document understanding (VDU) models: PICK [Yu et al. 2021], StrucTexT [Li et al. 2021], and DocFormer [Appalaraju et al. 2021]. These models combine deep learning and image processing, highlighting the use of synthetic data to improve performance. Carta

et al. 2024 explored information extraction without OCR using the document understanding Transformer (DONUT) model, employing a two-step training strategy with synthetic and real data. Performance improvement was observed when combining generated and real data.

Despite advancements in the field, the literature lacks comprehensive studies on deep models applied to Brazilian identity documents. This is partly due to the unavailability of real data for research purposes, as they contain sensitive private information. To address this issue, the Brazilian Identity Document (BID) Dataset [de Sá Soares et al. 2020] was released. It is a synthetic dataset containing three types of Brazilian identity documents, including the National Driver's License (CNH, from Portuguese *Carteira Nacional de Habilitação*), the focus of this study. All information in the BID dataset contains fake data, so the data presented in this paper does not expose any real person. The use of the BID Dataset ensures compliance with the General Data Protection Law (LGPD, from the Portuguese *Lei Geral de Proteção de Dados Pessoais*) [Planalto 2018].

The focus of this work is to evaluate the performance of pretrained deep models for text recognition of relevant typewritten fields in CNHs (i.e., excluding signatures and general texts) from the BID Dataset. We analyzed three size variations of TrOCR [Li et al. 2023], an open-source Transformer-based model that has been achieving state-of-the-art performance in several OCR tasks; Tesseract [Smith 2007], the most popular open-source OCR software in the literature; and GPT-4o, a proprietary large multimodal model based on the Transformer architecture. To evaluate the models' text recognition capabilities, we trained a detection model – as a secondary contribution– to locate relevant fields before recognition. Since the BID dataset does not explicitly link field data to its label (e.g., 759.844.427-79 $\rightarrow$ CPF), we implemented a post-processing algorithm to establish this mapping using predefined rules and heuristics.

Experimental results show that TrOCR Large achieved the top performance, with an F1-score of 96.33%. In summary, the main contributions of this work are: (i) a comparative study of deep learning models for text recognition of key fields in the Brazilian National Driver's License (CNH); a detection model to identify and locate relevant fields within the CNH; a semi-automatic approach for associating raw text annotations with their respective CNH fields, along with the processed annotations themselves; source code [1] to ensure the reproducibility of our results.

2. Material and Methods

This work evaluates models for text recognition in relevant fields of CNHs under two data extraction approaches: the detection-based approach, where relevant text entries (fields) are first located and then recognized using the investigated methods, and the zero-shot approach, where a prompt-based large multimodal model (GPT-4o and GPT-4o-mini) jointly performs field extraction and recognition. The following sections describe both approaches, as well as the the experimental setup used for performance evaluation.

2.1. Detection-based Approach

Figure 1 overviews the extraction pipeline for TrOCR models and Tesseract, comprising four stages: dataset preprocessing, field detection, text recognition, and performance

[1] `https://github.com/Suziane/OCR-CNH`

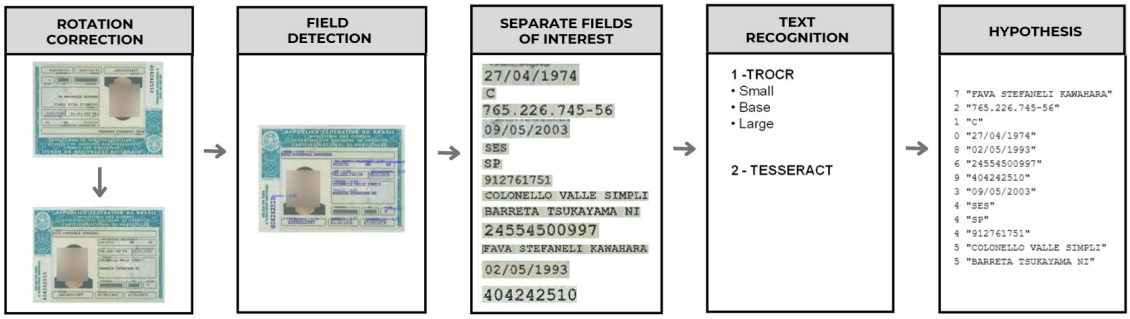

Figure 1. Processing pipeline, from receiving the image to the hypothesis generated by text recognition using method TROCR and Tesseract.

evaluation. The dataset is a BID subset with 3,600 front-face CNHs and annotated text entries (content and locations). During preprocessing, the orientation of the documents is corrected to address upside-down images. Additionally, a semi-automatic procedure is employed to identify and retain only the relevant text entries (fields of interest) from the annotations, ensuring proper alignment for subsequent tasks. The considered fields are (id - description): 0 - first driver's license; 1 - driving category; 2 - CPF; 3 - date of birth; 4 - identification document, issuer, and state; 5 - filiation; 6 - registration number; 7 - name of the driver; 8 - license validity; 9 - document mirror information.

In the following stage, a detection model is developed to locate the relevant fields. Part of the document samples (60%) are used to train the detection model, 20% to validation and the remaining 20% to detect the relevant fields, which are then used for text recognition and evaluation (test set). The following sections provide a more detailed discussion of the aforementioned stages.

2.1.1. Dataset Preprocessing

The BID collection includes images in four orientations: 0, 90, 180, and 270°. Orientation correction follows a simple strategy based on the position of the label (field id) NOME (name, in English) to ensure all images are set to 0°. In correctly oriented images, NOME appears in the upper-right quadrant of the image, while at 90, 180, and 270°, it shifts to the second, third, and fourth quadrants, respectively. The quadrant, and thus the rotation angle θ, is determined by comparing the tag's bounding box coordinates with the image dimensions. Correction is applied by rotating the image by $-\theta$.

The next step is identifying (labeling) the CNH fields. An annotation file, as shown in Figure 2 (left), contains raw text entries without explicitly linking field data to its label. In the example, the CPF field data (759.844.427-79) appears one line above its label. However, relative label-data positioning varies across annotation files and cannot be reliably used for field identification. To address this, we implemented an algorithm to label data fields and generate a new annotation file with consistent label-data alignment.

First, the annotation file is represented as a set of entries categorized into D-entries, corresponding to data field entries, and L-entries, corresponding to label entries, based on the categorization defined in Section 2.1. Each entry is represented as a tu-

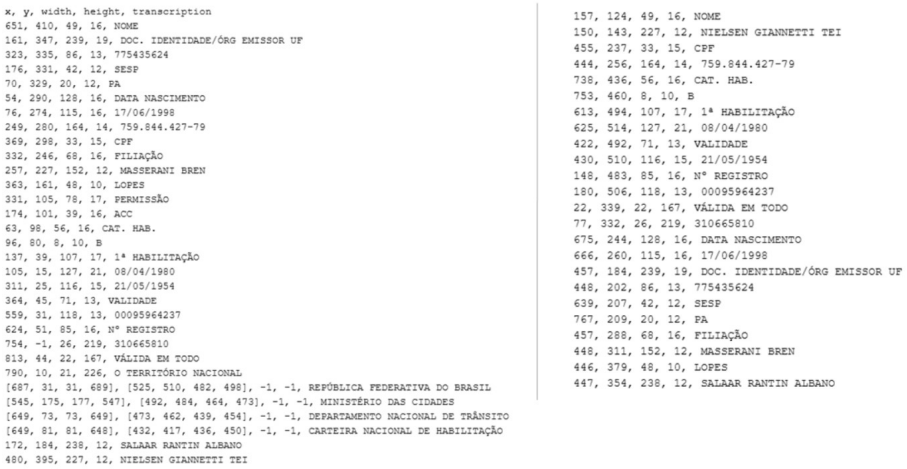

Figure 2. Original annotation file on the left and new file on the right (instance 00003615 of the BID Dataset).

ple $(pos, text)$, where pos denotes the top-left bounding box coordinate and $text$ is the transcribed content. Regular expressions were applied to identify data fields with unique formats, such as CPF, which follows the xxx.xxx.xxx-xx pattern. For this, D-entries were examined to detect matches between $text$ and the predefined regular expressions. The remaining data entries were identified through a one-to-one mapping between D-entries and L-entries. The mapping follows an iterative process, where an unlabeled D-entry is mapped to the closest L-entry positioned above the D-entry that has not been mapped yet. This process is repeated until all data entries have been appropriately matched. Validation rules were applied to ensure data integrity. For example, CAT. HAB. must contain a maximum of two letters and not correspond to a Brazilian state, and dates must follow the format (dd/mm/yyyy). The updated annotation file is shown in Figure 2 (right).

2.1.2. Relevant Fields Detection

The previous stage produced a dataset in which the images are correctly oriented and the text entries in the annotations have proper label (field) identification. From the overall dataset, 60% were used to train a model capable of detecting (locating and classifying) text entries corresponding to the fields of interest. The chosen model for this task is YOLOv8 [Jocher et al. 2023], a widely adopted variant of the YOLO (*You Only Look Once*) [Redmon 2016] family of architectures, recognized for balancing high accuracy with real-time detection capabilities. More specifically, we adopted the nano-sized variant, `yolov8n`, which provided accurate results while maintaining a manageable computational cost.

The YOLO model was trained for 40 epochs using the pretrained weights from the COCO dataset. During training, the batch size was set to 16, and the images were resized to 640 pixels. The learning rate was set to 0.01, with a momentum of 0.937 and a weight decay of 0.0005. A portion of 20% of the dataset (720 files) was reserved for validation, while another 20% was used for performance evaluation (further discussed in Section 2.3). Figure 3 shows an example of the annotation and the result of field detection.

Figure 3. Relevant field detection with YOLO: (left) CNH after orientation correction; (center) ground-truth bounding boxes (instance (00003604 from the BID Dataset); (right) field detection result (instance 00004155 from the BID Dataset).

Field detection was integrated into the pipeline to enable practical applications and testing with other documents, although this work focuses on text recognition.

2.1.3. Text Recognition via OCR

At this stage, the detected text entries should undergo text recognition. Two distinct OCR methods were evaluated: TrOCR and Tesseract. Both methods rely on neural networks to recognize text from images, but they differ in their underlying architectures and training strategies. There was used pretrained models for both methods, without retraining. The results of the OCR methods were saved in a structured format, associating each recognized text with its corresponding field for further evaluation. A brief description of the two methods is provided in the following.

TrOCR TrOCR [Li et al. 2023] is a Transformer-based OCR model designed for end-to-end text recognition. The model integrates a visual encoder (BEiT [Bao et al. 2021] or DEiT [Touvron et al. 2021]) with a language decoder (MiniLM [Wang et al. 2020] or RoBERTa [Liu et al. 2019]) responsible for generating text. TrOCR operates by dividing images into 16×16 pixel patches, converting them into linear representations, and then using the Transformer encoder to process the image. The text decoder then generates tokens in an autoregressive fashion for accurate transcription. In this study, we evaluated three variations of TrOCR: $\text{TrOCR}_{\text{SMALL}}$ (62M parameters), which integrates the encoder of $\text{DeiT}_{\text{SMALL}}$ with the decoder of MiniLM; $\text{TrOCR}_{\text{BASE}}$ (334M parameters), which combines the encoder of $\text{BEiT}_{\text{BASE}}$ with the decoder of $\text{RoBERTa}_{\text{LARGE}}$; and $\text{TrOCR}_{\text{LARGE}}$ (558M parameters), which pairs the encoder of $\text{BEiT}_{\text{LARGE}}$ with the decoder of $\text{RoBERTa}_{\text{LARGE}}$. Each variation was fine-tuned on the SROIE dataset [Huang et al. 2021], a benchmark for extracting key information from scanned receipts, to optimize recognition of printed documents.

Tesseract OCR Although the focus of this research was initially on evaluating TrOCR, we also included Tesseract [Smith 2007], a widely recognized open-source OCR engine. Tesseract was originally developed by Hewlett-Packard (HP) between 1984 and 1994 and was later acquired by Google in 2006. Since then, it has been maintained as an open-source project. Tesseract supports multiple languages and provides high-quality text recognition. The version used in this study is Tesseract 4c, which is based on the

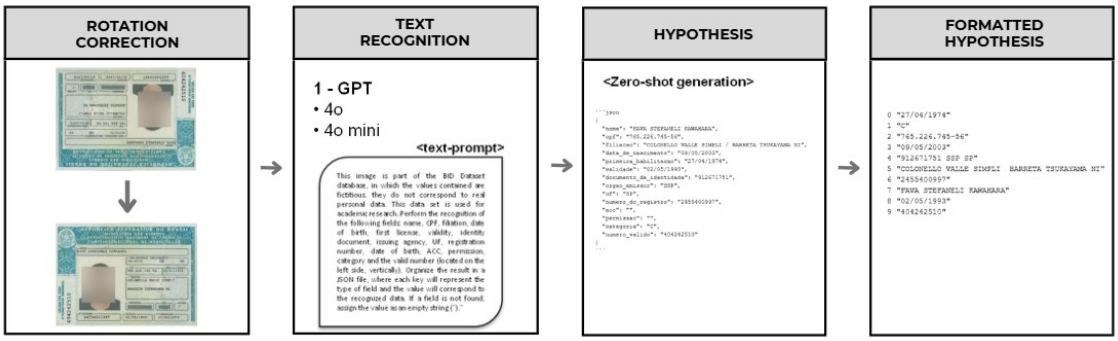

Figure 4. Processing pipeline, from receiving the image to the hypothesis generated by text recognition using method GPT.

Long Short-Term Memory (LSTM) architecture [Hochreiter and Schmidhuber 1997], a type of recurrent neural network that has proven highly effective for sequence modeling tasks. Tesseract was used with the option `--psm 7`, indicating that the image would be treated as a single line of text. The parsing language was set to Portuguese using `(lang="por")`.

2.2. Zero-shot Recognition Approach

In this approach, extraction of relevant fields and text recognition is performed in a single inference by the OpenAI GPT-4 Omni (GPT-4o)[2], a large multimodal model based on the GPT (Generative Pretrained Transformer) architecture [Radford et al. 2018]. This model combines the strengths of GPT's natural language understanding with advanced capabilities for handling visual data, making it suitable for tasks such as document analysis and text recognition in structured documents. While there has been speculation that the GPT-4o model contains trillions of parameters, OpenAI has not officially disclosed its exact size. There is a cost associated with using the API, typically calculated per token (costs are discussed in Section 3).

In practice, GPT (GPT-4o and GPT-4o-mini) is accessed via API by passing a textual prompt (e.g., asking the model to extract specific fields from a document) along with a CNH image as illustrated in Figure 4. The process occurs sequentially: first, the image is adjusted to correct its rotation, then it is sent to the API. The model responds with a JSON file containing the results, which, in turn, are formatted to be used in the metrics and evaluation steps. The exact prompt used in this work was: *"This image is part of the BID Dataset database, in which the values contained are fictitious, that is, they do not correspond to real personal data. This data set is used for academic research. Perform the recognition of the following fields: name, CPF, filiation, date of birth, first license, validity, identity document, issuing agency, UF, registration number, date of birth, ACC, permission, category and the valid number (located on the left side, vertically). Organize the result in a JSON file, where each key will represent the type of field and the value will correspond to the recognized data. If a field is not found, assign the value as an empty string ('')."* Each sample from the test set (20% of the dataset, as mentioned in Section 2.1) was submitted to the API for inference, and the results were saved for performance evaluation and comparison with the detection-based approach.

[2] `https://openai.com/index/hello-gpt-4o.`

2.3. Performance Evaluation

Performance was evaluated using both word- and character-based metrics. Word-based evaluation provides a broader perspective of performance at the document level, while character-based evaluation enables a more granular analysis at the field level.

2.3.1. Word-based Evaluation

Word-based analysis, as utilized by the authors of TrOCR to evaluate performance on scanned receipts from the SROIE dataset [Huang et al. 2021], relies on the computation of word-level precision, recall, and F1-score:

$$\text{Precision} = \frac{\text{Correct matches}}{\text{Number of detected words}}, \tag{1}$$

$$\text{Recall} = \frac{\text{Correct matches}}{\text{Number of ground truth words}}, \quad \text{and} \tag{2}$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

This approach is particularly suitable for the fields filiation (ID-5) and name of the driver (ID-7), which are expected to contain more extensive text content. Given that the mentioned metrics are highly sensitive to a low number of words, the text from fields ID-5 and ID-7 within each test sample was concatenated. Subsequently, the resulting strings were further concatenated across all documents, producing a single string for each recognition method. The same procedure was applied to the ground-truth annotations, enabling the computation of the evaluation metrics. It is worth noting that the strings are treated, for the purpose of word-level metric computation, as bags of words rather than sequences, i.e., the order of words is not considered relevant. This provides a general and approximate view of the performance of the different methods. Although the WER metric is widely used in text recognition tasks, we did not apply it in our investigation, as it is better suited for documents with more continuous text. This decision is based on the dataset used, which predominantly consists of short text segments, single words, or dates. WER can be misleading for short texts, as small errors have a large impact.

2.3.2. Character-based Evaluation

The character-based evaluation relies on the *Character Error Rate* (CER) [Neudecker et al. 2021] metric, widely used for evaluation of OCR systems. The CER represents the fraction of characters incorrectly predicted by the model: the lower the value, the better the performance, with a value of 0 indicating perfect recognition. Formally,

$$\text{CER} = \frac{i + s + d}{n}, \tag{4}$$

Method	Precision (%)	Recall (%)	F1 (%)
TrOCR Small	87.367	87.235	87.301
TrOCR Base	93.343	93.229	93.286
TrOCR Large	**96.391**	**96.259**	**96.325**
GPT-4o	87.676	87.028	87.351
GPT-4o-mini	85.781	80.358	82.981
Tesseract	72.408	72.833	72.620

Table 1. Precision, Recall and F1-Score considering the concatenation of ID 5 (filiation) and ID 7 (name of the driver).

where n is the total number of characters in the ground-truth reference text, and i, s, and d represent the minimal number of **insertions**, **substitutions**, and **deletions**, respectively, required to transform the ground-truth reference text into the OCR output. CER is computed field-wise for each document sample in the test set and subsequently averaged across all samples. To provide a comprehensive view of the overall performance, the mean and standard deviation of the CER values are reported. This analysis is particularly beneficial for fields containing information with non-alphabetic symbols, such as CPF, date of birth, and registration number, where character-level accuracy is essential.

2.4. Computational Environment

Experiments were conducted on Google Colab PRO, equipped with an NVIDIA Tesla T4 GPU and 16 GB of VRAM. The JiWER [Jitsi 2024] package was used for CER computation, while word-level evaluation employed an algorithm adapted from Microsoft's implementation (`https://github.com/microsoft/unilm/blob/master/trocr/scoring.py`).

3. Results and Discussion

The results in Table 1 correspond to the evaluation of the set of all concatenated names and surnames, including both the filiation (ID 5) and the name of the driver (ID 7). It can be observed that the TrOCR Large model obtained the best overall performance, reaching 96.391% precision, 96.259% recall, and 96.325% F1-score, reflecting a high capacity to correctly identify information and minimize errors. The Tesseracts model, which presented a precision of 72.408%, had lower performance in the three metrics compared to the other models, indicating limitations. Models such as GPT (GPT-4o) and GPT Mini (GPT-4o-mini) demonstrated intermediate results.

According to Figure 5, it is possible to analyze the CER and performance of different models in the text recognition task for each ID. Since insertions are theoretically unlimited, CER values greater than 1 are mathematically possible and usually show severe errors such as excessive insertions, deletions, or substitutions. Furthermore, these values tend to occur when the reference text is very short, since even a few errors disproportionately inflate the ratio. The top right list indicates the number of test instances successfully processed. Failures, in this context, were observed only for GPT-4o, GPT-4o-mini, and Tesseract. Tesseract processes the fields individually but failed to recognize some of them. In contrast, GPT-4o and GPT-4o-mini received the complete document image, being responsible for identifying the field and recognize the text content, a harder task that may

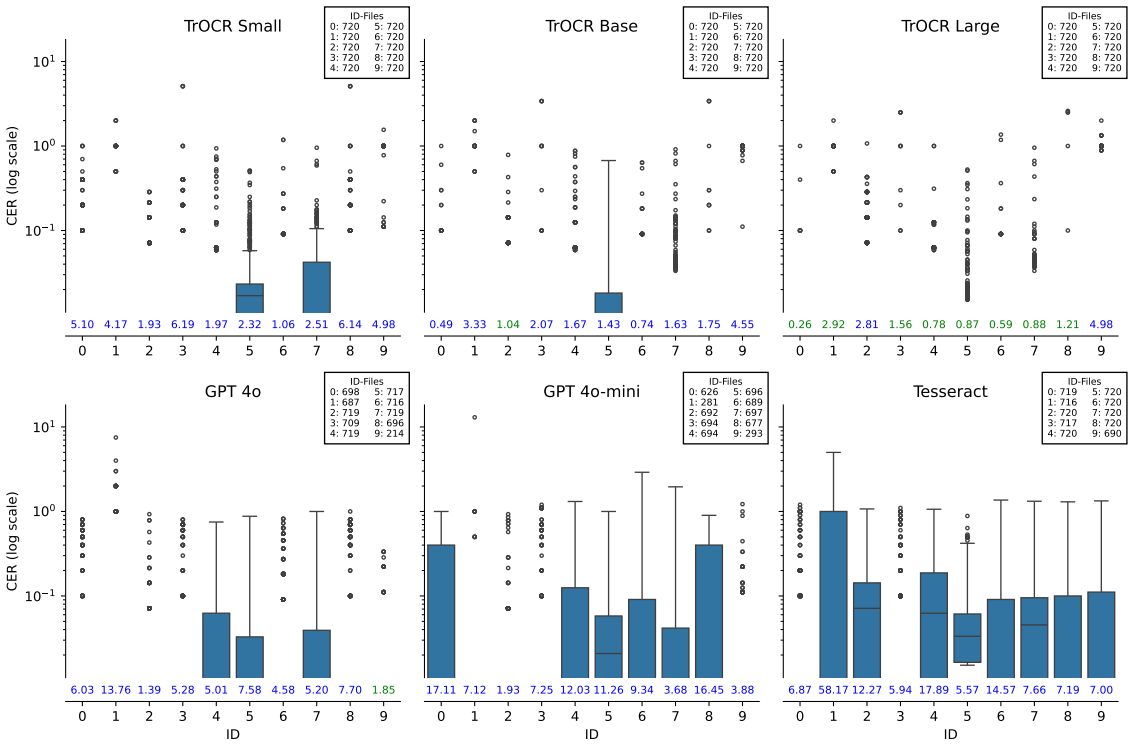

Figure 5. CER distribution for individual IDs. The blue values above the x-axis indicate the average CER (%), while the numbers inside the boxes represent successfully processed instances. Values in green indicate the best overall performance across the models. ID assignments are as follows: 0 – first driver's license, 1 – driving category, 2 – CPF, 3 – date of birth, 4 – identification document, issuer, and state, 5 – filiation, 6 – registration number, 7 – driver's name, 8 – license validity, and 9 – document mirror information.

explain the higher number of failures. The TrOCR models stand out for their accuracy and consistency, especially the Base and Large models, which present significantly low CERs accompanied by smaller standard deviations compared to the other models. TrOCR Large is yielded the best overall performance, with CER values ranging from 0.26 (ID 0 - first driver's license) to 4.98 (ID 9 - document mirror), reflecting smaller scatter in the results, stability, and accuracy, even on challenging IDs. TrOCR Base also presents interesting results, such as the CER of 1.04 on ID 2 (CPF). TrOCR Small achieved the highest CER rates comparing to it counterparts, but outperformed the Tesseract and GPT models in most fields.

The GPT models perform at an intermediate level, with higher error rates and greater variability compared to TrOCR. As expected, GPT-4o is more consistent than GPT-4o-mini, with CER values from 1.39 in ID 2 (CPF) to 13.76 in ID 1 (driving category), its most challenging case. Its best overall performance is observed in ID 9 (document mirror), with a CER of 1.85. GPT-4o-mini yielded CER values from 1.93 in ID 2 (CPF) to 17.11 in ID 0 (first driver's license). While it achieves reasonable performance in ID 2, close to that of GPT-4o, its higher variability and weaker results in complex IDs, like ID 0, indicate less robustness. Tesseract demonstrates the most inconsistent performance among the evaluated methods. Its CER values vary widely, ranging from 5.57 (ID

5 - filiation) to 58.17 (ID 1 - driving category), reflecting significant challenges in maintaining accuracy across different scenarios. Tesseract performs particularly poorly in the ID 1 (driving category) test, with a CER of 58.17%. This ID corresponds to the category field, which contains few characters, likely contributing to the model's reduced accuracy. Although Tesseract recovers somewhat in other IDs, it does not reach the accuracy levels of the TrOCR or GPT models, underscoring its limitations compared to more advanced OCR solutions.

Time and Financial Costs The TrOCR models achieved the best accuracy but at a significant processing time cost. In an experiment with 10 document samples, TrOCR Large required 29 minutes and 43 seconds, TrOCR Base took 14 minutes and 46 seconds, and TrOCR Small completed the task in 4 minutes and 4 seconds. In contrast, Tesseract, though less accurate, processed the same samples in just 1 minute and 6 seconds. The processing cost per image was approximately $0.0018 for GPT-4o and $0.0013 for GPT-4o-mini, calculated using the OpenAI Tokenizer (`https://platform.openai.com/tokenizer`) and Pricing (`https://openai.com/api/pricing`) platforms. The total cost for all GPT-4o and GPT-4o-mini experiments was approximately $4.81, based on November 2024 reference values.

4. Conclusion and Future Work

This study investigated the performance of different text recognition models applied to reading driver's licenses, with the aim of identifying the most suitable ones in terms of accuracy, reliability and robustness. The analysis provided insights into the capabilities of these models in extracting textual information in specific fields.

The comparative analysis indicated the TrOCR model, especially in its Large variant, as the most accurate and robust for OCR tasks, standing out even without the need for retraining. Although it requires an additional step for field detection, TrOCR proved to be efficient, with consistent performance and being an open-source solution. GPT-based models also performed well, but with greater variability in results, limiting their reliability in some scenarios. On the other hand, Tesseract revealed difficulties in dealing with complex data, being less suitable for high-precision demands.

Despite its contributions, the study has limitations, such as the exclusive focus on driver's licenses and the lack of analysis of geometric distortions or adverse conditions. These limitations restrict the generalization of the results to other types of documents and application scenarios. As future directions, we suggest expanding the scope of the analysis to include other documents, such as passports or IDs, and evaluating the impact of geometric distortions, such as rotations and tilts. These advances can increase the practical applicability of the models and offer more robust solutions to real-world challenges in text recognition systems.

Acknowledgements. The authors would like to thank FAPES/UnAC (No. FAPES 1228/2022 P 2022-CD0RQ, No. SIAFEM 2022-CD0RQ) for the financial support provided through the UniversidaES System.

References

Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., and Manmatha, R. (2021). Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.

Attivissimo, F., Giaquinto, N., Scarpetta, M., and Spadavecchia, M. (2019). An automatic reader of identity documents. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3525–3530. IEEE.

Bao, H., Dong, L., Piao, S., and Wei, F. (2021). BEiT: BERT Pre-Training of Image Transformers. *arXiv preprint arXiv:2106.08254*.

Baviskar, D., Ahirrao, S., Potdar, V., and Kotecha, K. (2021). Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. *IEEE Access*, 9:72894–72936.

Carta, S., Giuliani, A., Piano, L., and Tiddia, S. G. (2024). An end-to-end ocr-free solution for identity document information extraction. *Procedia Computer Science*, 246:453–462. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

Castelblanco, A., Solano, J., Lopez, C., Rivera, E., Tengana, L., and Ochoa, M. (2020). Machine learning techniques for identity document verification in uncontrolled environments: A case study. In *Pattern Recognition*, pages 271–281, Cham. Springer International Publishing.

Chandra, A. and Stefanus, R. (2021). An end-to-end optical character recognition pipeline for indonesian identity card. In *2021 9th International Conference on Information and Communication Technology (ICoICT)*, pages 307–312.

de Sá Soares, A., das Neves Junior, R. B., and Bezerra, B. L. D. (2020). BID Dataset: a challenge dataset for document processing tasks. In *Anais Estendidos do XXXIII Conference on Graphics, Patterns and Images*, pages 143–146. SBC.

Gov, A. (2024). E-commerce no brasil cresce 4% e alcança r$ 196 bi em 2023. `https://agenciagov.ebc.com.br/noticias/202409/e-commerce-no-brasil-cresce-4-e-alcanca-r-196-bi-em-2023` Accessed: December 9, 2024.

Hoai, D. P. V., Duong, H.-T., and Hoang, V. T. (2021). Text recognition for vietnamese identity card based on deep features network. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(2):123–131.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., and Jawahar, C. V. (2021). ICDAR2019 competition on scanned receipt OCR and information extraction. *CoRR*, abs/2103.10213.

Jitsi (2024). Jiwer: A python package for word error rate and character error rate computation. `https://jitsi.github.io/jiwer/` Accessed: December 6, 2024.

Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics yolov8. `https://github.com/ultralytics/ultralytics` Accessed: November 20, 2024.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. (2023). Trocr: Transformer-based optical character recognition with pre-trained models. In *The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)*.

Li, Y., Qian, Y., Yu, Y., Qin, X., Zhang, C., Liu, Y., Yao, K., Han, J., Liu, J., and Ding, E. (2021). Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1912–1920.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 364.

Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A., and Pletschacher, S. (2021). A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 13–18, New York, NY, USA. Association for Computing Machinery.

Planalto (2018). Lei nº 13.709, de 14 de agosto de 2018. `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm` Accessed: November 25, 2024.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Redmon, J. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Smith, R. (2007). An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Subramani, N., Matton, A., Greaves, M., and Lam, A. (2021). A survey of deep learning approaches for ocr and document understanding.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Wojcik, L., Coelho, L., Granada, R., Führ, G., and Menotti, D. (2023). NBID dataset: Towards robust information extraction in official documents. In *2023 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 145–150.

Yu, W., Lu, N., Qi, X., Gong, P., and Xiao, R. (2021). Pick: processing key information extraction from documents using improved graph learning-convolutional networks. In *2020 25th International conference on pattern recognition (ICPR)*, pages 4363–4370. IEEE.