Acta Scientiarum
http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8672
Doi: 10.4025/actascitechnol.v47i1.69814

Optical character recognition system with natural language processing for data recovery on scanned old academic card reports

Edwin Casas-Huamanta[1*], Lloy Pinedo[2], Enrique Barbachán-Ruales[3], Ángel Cárdenas-García[2], Luis Rossel-Bernedo[4] and José Seijas-Díaz[5]

[1]Escuela de Educación Superior Pedagógica Pública Monseñor Elías Olázar, Yurimaguas, Alto Amazonas, 16501, Loreto, Peru. [2]Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional de San Martín, Tarapoto, 22200, San Martín, Peru. [3]Facultad de Tecnología, Universidad Nacional de Educación Enrique Guzmán y Valle, Lurigancho, 15472, Chosica, Perú. [4]Facultad de Ingeniería Estadística e Informática, Universidad Nacional del Altiplano, Puno, 21001, Perú. [5]Facultad de Ciencias Administrativa y Contables, Universidad Nacional Autónoma de Alto Amazonas, Yurimaguas, Alto Amazonas, 16501, Loreto, Peru. *Author for correspondence. E-mail: rcasas@unaaa.edu.pe

ABSTRACT. In the digital age, preserving and effectively retrieving historical academic records has a significant challenge, especially when these documents only exist in deteriorated physical formats. We propose an approach to recover data from scanned documents of grade records, by using image processing and Natural Language Processing (NLP) to enhance the accuracy of Optical Character Recognition (OCR) in these documents, essential for the preservation of digital records. Our three-step methodology: first, improves the quality of the scanned image; then, extracts text using OCR and NLP techniques to retrieve data from old physical grade cards; and finally, the extracted data is corrected using Chat-GPT and prepared for upload. The results are assuring, showing an impressive Character Error Rate (CER) of 2.15% and a Word Error Rate (WER) of 7.05%, demonstrating the high accuracy of the OCR system used and its ability to precisely extract text from scanned documents. These low error rates achieved, as a result to the successful implementation of pre-processing and post-processing techniques, as well as the use of an advanced OCR tool, underscore the potential of this OCR approach to effectively extract information from documents.

Keywords: accuracy; image processing; chat-GPT; digital records; preservation.

Received on October **2**, 2023.
Accepted on March 15 2024.

Introduction

In the field of documentary preservation, the physical deterioration of records presents a significant challenge (Ríos Martínez, 2021). While it is feasible to store contemporary documentation in digital repositories, the effective preservation of historical documents, generated prior to the digital era, poses unique dilemmas (Ahmad & Rafiq, 2023; Barrueco & Termens, 2022; Kraus et al., 2021). The manual transcription of these records into digital formats is a laborious and often economically inefficient task. In this context, digitization has emerged as a prevalent solution (Nadkarni & Prügl, 2021).

However, digitized versions do not always faithfully replicate the functionality of the original physical documents. Their modification and editing can be restrictive, and often, these reproductions lack optimal quality. This is where Optical Character Recognition (OCR) plays a leading role. Recent innovations have enhanced its capacity, allowing the conversion of vast sets of scanned images into editable text (Ghosh et al., 2022; Kashinath, Jain, Agrawal, Anand, & Singh, 2022). Additionally, the incorporation of machine learning and deep learning techniques in this domain promises optimization in the post-processing of OCR output, minimizing discrepancies in character recognition (Ahmed et al., 2023; Lombardi & Marinai, 2020).

The quality improvement of digitized documents through advanced image processing techniques and Natural Language Processing (NLP) opens horizons for a series of essential applications (Mah, Skalna, & Muzam, 2022). Such improvement not only enables access to historical repositories, but empowering researchers with contextualized information but also plays a vital role in preserving significant records for future generations, safe them from the inevitable time damage (Kang, Cai, Tan, Huang, & Liu, 2020). Additionally, this enhancement in quality provides organizations with more robust tools for the management and regulatory compliance of their records (Zanabria-Ortega, 2022).

Contrasting with the inherent challenge of preserving old documents, the imperative transition towards digitizing physical records is further accentuated in the face of the exponentially growing volume of data characteristic of contemporary societies. Traditional methodologies of manual data entry are not only tedious and labor-intensive but are also fraught with potential errors, whose repercussions can be particularly catastrophic in sensitive areas such as legal or medical records (Kim, Choi, Park, & Kim, 2022). In light of this, techniques for automated data extraction are gaining prominence and have captured substantial interest from both the academic and industrial realms (Memon, Sami, Khan, & Uddin, 2020).

OCR represents a well-established technique for automatic data extraction that focuses on discerning text, whether printed or handwritten, derived from images or digitized documents. Despite OCR's progressive adoption over various decades, intrinsic challenges persist in the accurate recognition of text from damaged or suboptimal images (Sulaiman, Omar, & Nasrudin, 2019). These challenges encompass factors such as inconsistent lighting, noisy interferences, distortions, and typographic diversity. Consequently, elevating the precision and robustness of OCR has outlined a primary focus in contemporary research in computer vision and image processing (Zeng et al., 2023).

Given the constraints of conventional OCR paradigms, current inquiries have shifted their focus towards the amalgamation of machine learning and deep learning techniques to enhance OCR efficiency. Among these techniques, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-type architectures emerge notably (Kohli, Agarwal, & Kumar, 2022; Liu, Wang, & Shi, 2023). CNNs have proven apt for extracting intrinsic features from images, while RNNs excel in modeling sequential textual information (Rahali & Akhloufi, 2023). On the other hand, Transformer models, with BERT as a prominent example, have established benchmarks in natural language processing tasks, finding promising adaptations in the OCR domain (Patwardhan, Marrone, & Sansone, 2023).

In this paper, we propose a novel approach that combines image processing techniques and NLP to enhance OCR performance on scanned documents of student grade cards. Our approach involves preprocessing the scanned images to improve image quality, followed by the use of a deep learning-based OCR model that incorporates NLP techniques to improve recognition accuracy. We also evaluate our approach on a large dataset of grade card images and compare it with several state-of-the-art OCR methods.

Material and methods

Our proposed approach focuses on retrieving information from archived former texts, which have been subject to wear over time, causing damage to the paper or fading of the ink. Since not all documents deteriorate at the same rate, analyzing information from such documents can be challenging. Therefore, our three-step procedure aims to address these challenges.

In the first step, we digitally scan former physical grade slips (paper documents used by educational institutions to manually record student grades and evaluations), using high-quality scanners (with A4 scanning format capabilities, 1200 x 1200 dpi resolution capacity, and scanning modes of 24-Bit Color, 8-Bit Grayscale, 1-Bit Monochrome). Then, we apply image-preprocessing techniques to improve the quality of the scanned images, including "De-speckle" to reduce grayscale artifacts in images, "De-skew" to correct the orientation of the text, and binarization methods to enhance the clarity of the text in the images. These techniques ensure that the scanned images are of high quality before moving on to the next step.

In the second step, we employ OCR and NLP techniques to extract data from the old physical report cards. OCR is particularly useful for converting scanned images into text, while NLP techniques can be used to filter out irrelevant information, such as names and indications, which are not part of the English dictionary. Subsequently, the extracted data is analyzed for errors, which are corrected using RoBERT to predict and amend misspelled and inaccurately transcribed words. This process has been shown to significantly reduce the Word Error Rate (WER) and the Character Error Rate (CER).

In the final step, the extracted data is prepared for uploading into a database or another storage format. This involves organizing the data according to predefined categories, such as student names, identification numbers, and grades, ensuring the data is error-free. This step is essential to ensure that the extracted data is easily accessible and can be used for further analysis.

Overall, our approach offers a comprehensive solution for retrieving information from ancient archived texts. By combining image-preprocessing techniques, OCR, and NLP, we can overcome the limitations of traditional OCR methods and extract high-quality data from worn physical documents.

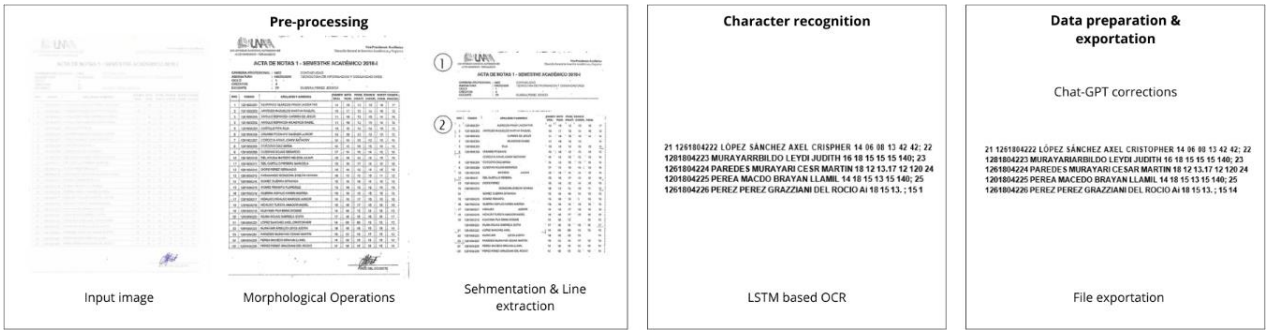


Figure 1. Steps in the OCR pipeline.

Figure 1 illustrates the primary steps proposed for retrieving data from card reports: I. Pre-processing; II. Character Recognition; III. Data preparation and export.

Pre-processing: this initial step in the OCR pipeline focuses on enhancing the image quality of the input document. The aim of improvements in this phase is to remove noise and artifacts from the image, which can adversely affect OCR accuracy. A crucial aspect of pre-processing is segmentation. As previously noted, segmenting the images is a beneficial method to process noisy documents. For our study, the documents in our dataset were segmented into two categories: the tables within the documents, and the titles along with metadata.

Character recognition: this involves transforming images of typed, handwritten, or printed text into machine-encoded text. Enhancements in this step aim to augment both the speed and accuracy of OCR. Any inaccuracies in the OCR process could result in omitted or erroneous text in the output. We apply deep learning techniques to execute this phase.

Data preparation and export: post-processing ensues after OCR. During this phase, further processing of the extracted text is done to rectify errors and enhance the output's accuracy. The objective of improvements here is to amend any inconsistencies or errors in the text, such as misspellings, grammatical mistakes, and formatting challenges. Upon completing the data preparation, the concluding step in the OCR pipeline involves exporting the extracted text in an appropriate format, such as a text file or a searchable PDF. The goal for enhancements in this step is to guarantee that the exported data is both accurate and in the intended format since discrepancies in the export, process might lead to unusable or imprecise data.

Each of these steps is described in further detail below.

Pre-processing

To retrieve information from noisy scanned PDF documents, it is pivotal to convert these documents into a more adaptable format. To achieve this, we transformed all documents into the JPG format using the CV2 Python library. Following this conversion, we engaged in image processing to rejuvenate the data. To streamline the analysis and data reconstruction, it was imperative to bifurcate the images into two distinct categories: tables and titles coupled with metadata. This distinction was crucial as the processing tailored for one segment could impede the OCR performance of the other. To circumvent this, we initiated segmentation in a novel color space.

Image processing for the titles and metadata commenced with an image threshold set between 10 and 199. This step was instrumental in purging the image of stains and assorted noise prevalent in the scanned document. Subsequently, the image values were inverted, and a morphological open operation was executed on the outcome. Concluding this phase, the image values were reverted to recapture the original image orientation, culminating in the successful reconstruction of the metadata.

Although the tables underwent a processing sequence akin to that of metadata, an extra step was incorporated to erase lines. This facilitated smoother data extraction via OCR. Nevertheless, this line removal proved detrimental for the titles and metadata, distorting vital information and diminishing OCR precision. This underscored the need for our prior segmentation. For the treatment of lines within tables, a Gaussian filter was employed, succeeded by a border enhancement for both lines and text. This result was preserved as a duplicate. Leveraging a morphological Black-hat kernel from CV2, the differences were extracted using cv2.morphologyEx. With these distinct line images in hand, information was siphoned off, completing the image processing for the tables.

In essence, to revitalize information from grainy scanned PDFs, the documents were transmuted to the JPG format and bifurcated into two categories: tables and titles paired with metadata. The image-processing

trajectory encompassed myriad steps, including image thresholding, inversion, morphological open operation, Gaussian filtering, and border enhancement. Each step was meticulously crafted to cater to the unique challenges of each segment and to bolster OCR accuracy.

Character recognition

In the endeavor to analyze the reinstated information from scanned documents, a pivotal subsequent step post-image preprocessing involves the utilization of Optical Character Recognition (OCR) methodologies. Specifically, our research employed Pytesseract, a renowned open-source OCR tool, to discern and retrieve characters from the refined images. Notably, the product of the OCR endeavor is a textual string, which, upon further scrutiny using Natural Language Processing (NLP) techniques, can yield salient information. Techniques such as named entity recognition, text classification, and sentiment analysis are invaluable in distilling insights and patterns from the resultant data.

It is paramount to acknowledge that while OCR stands as a transformative technology, it is not without its imperfections. The recognition phase occasionally witnesses inaccuracies. Hence, it is indispensable to conduct rigorous quality audits and amend the extracted content to guarantee its precision and comprehensiveness. Enhancing OCR output quality can be achieved through the judicious calibration of preprocessing and OCR parameters, incorporating methodologies like adaptive thresholding, deblurring, and strategic character segmentation.

In essence, merge of image preprocessing, OCR, and NLP offers a robust strategy for renewing and dissecting information from grainy scanned PDF artifacts. Harnessing this suite of techniques, we can unearth invaluable insights from vast repositories of unstructured content, which can be instrumental in shaping decisions and fostering innovation across varied sectors.

Upon finalizing image preprocessing and meticulous segmentation, our trajectory necessitated the extraction of data from the renewed documents through OCR methodologies. In this endeavor, Pytesseract was our tool of choice. Its reputation for accuracy and user-friendliness, underpinned by training on an extensive dataset encompassing a myriad of fonts, dialects, and image resolutions, makes it an unrivaled asset for OCR undertakings.

Our method involved feeding the preprocessed images into Pytesseract. This engine meticulously identifies and retrieves the embedded textual content. The resultant output—a concatenated string—encapsulates the recovered data, which can be migrated to diverse formats to facilitate subsequent examinations or applications. Such extracted data lays the groundwork for insights, facilitating endeavors like statistical inferences, visual data representations, or sophisticated natural language processing.

Delving into the mechanics, Pytesseract harnesses an optical character recognition system predicated on a neural network. This neural framework is modeled on a convolutional recurrent architecture—an exemplary model for sequence-based prediction tasks, such as OCR. The network undergoes training on an expansive corpus of textual imagery, empowering it to identify a plethora of text styles, sizes, and fonts.

Ingesting an image, this network deploys multiple convolutional and recurrent layers to distill features and encode the text's contextual essence. The terminal output is a nuanced sequence of character probabilities corresponding to each image position. Pytesseract adopts a beam search decoding methodology to pinpoint the most probable character sequence based on these output probabilities. This decoder inherently factors in adjacent characters' contexts to enhance the resultant OCR's accuracy.

In addition, synergizing image processing techniques with OCR—anchored in Pytesseract—proffers a potent mechanism for recovering data from noise-laden scanned PDFs. Initial stages necessitate document format conversion and meticulous image segmentation. Consequent stages employ intricate image processing techniques like thresholding and morphological interventions to siphon off relevant data. Finally, with the OCR process orchestrated by Pytesseract, textual content is extracted and prepared for multifaceted applications.

Data preparation and exportation

Data preparation and exportation is a crucial step in any data-driven project, as it ensures that the extracted data is accurate, consistent, and in a suitable format for further analysis or use. In the context of OCR, data preparation and exportation involve post-processing the extracted text to correct errors and inconsistencies, and exporting it in a desired format such as a text file or a searchable PDF. In this section, we will discuss the data preparation and exportation process used in our OCR pipeline, highlighting the methods and tools used to ensure that the extracted data is of high quality and suitable for use in downstream applications.

Data preparation

Once the OCR was completed, the string chains often had issues with space detection at specific points in the converted text. To address this problem, the string method incorporated in Python 3 was used. This was helpful in eliminating the extra spaces and undetected punctuation marks, as well as formatting the tables and metadata in the strings, making them ready to be exported as information to a Pandas data frame.

In addition, there were some cases where the OCR did not recognize certain characters correctly, leading to errors in the output data. To overcome this, manual correction was required to ensure the accuracy of the data. This involved going through the text data and identifying any mistakes or errors, and then manually correcting them.

After the data was corrected and formatted, post-processing techniques were used to analyze and extract meaningful insights from the data. This involved techniques such as data cleaning, filtering, and visualization, to identify patterns and trends in the data. These techniques helped to further refine the data and extract valuable insights that could be used for various applications, such as business intelligence or data analytics.

Data exportation

Once the post-processing was complete, the final step was to convert the string chains into data frames, which could be used to create a database. The data frames were formatted appropriately to ensure that the information was properly structured and organized. Finally, to complete the analysis, the data was exported into a .CSV format with the same name system as the original PDF files.

This allowed the data to be easily shared and analyzed in other software or platforms. The resulting .CSV files contained all the necessary information in a clean and organized format. This process made it easier to search, sort and filter the data, enabling efficient analysis of the scanned documents.

Exporting the data in a CSV format is useful because it allows for easy integration of the analyzed data into other software applications or databases. CSV is a commonly used file format that can be easily imported into popular data analysis tools such as Microsoft Excel, R, or Python.

Once the data is in a CSV format, it can be easily shared with others who may not have access to the original scanned PDF documents, making it more accessible and widely available. Additionally, having the data in a structured format such as CSV allows for easier manipulation and analysis of the data using various data analysis tools and techniques.

The post-processing stage was a critical component of our data analysis process, as it helped ensure that the data, we analyzed was accurate, consistent, and usable. This stage involved a combination of manual and automated techniques, including text cleaning and normalization, data filtering and validation, and outlier detection and removal. By carefully processing the data in this way, we were able to correct errors and inconsistencies in the text, remove unnecessary or irrelevant data, and ensure that the data we analyzed was representative and reliable.

Once the data had been thoroughly processed and cleaned, the next step was to export it in a format that could be easily accessed, shared, and analyzed. We chose to export our analyzed data in CSV format, which offers a range of benefits over other file formats. CSV files are widely used and supported across a variety of software platforms, making it easy to import and export data between different systems. Additionally, CSV files are easy to read and interpret, making them an ideal choice for sharing data with others, whether they are fellow researchers, stakeholders, or members of the public.

Furthermore, exporting our analyzed data in CSV format allowed us to take advantage of a range of tools and techniques for data analysis, including data visualization and machine learning. By exporting our data in this way, we were able to leverage the power of tools like Python and R to generate insights, identify patterns and trends, and make predictions based on the data. This flexibility and interoperability made the process of analyzing our data much more efficient and effective, and allowed us to derive more meaningful and actionable insights from it.

Results and discussion

Titles and metadata

Figure 2 shows the comparison between the changes in the images after each filter and morphological operation. To effectively prepare images for OCR analysis, we systematically process them through a series of fundamental steps. Initially, we transform the color images into grayscale, simplifying them and enhancing

their processability. Subsequently, we implement an adaptive threshold, segmenting the image into different black and white regions. This segmentation is achieved by evaluating the pixel values within a localized neighborhood around each pixel, then applying a threshold based on the average intensity. To make the binary result more readable for OCR, we invert the image, setting the background to white and the text to black.

In our quest to refine the image further, we incorporated morphological operations, which are techniques crafted to modify the contour of objects within an image. We utilized an opening operation to counteract minor noise and refine the character edges. Sequentially, a closing operation was executed to bridge gaps and mend fragmented lines. The granularity of these operations was dictated by a structuring element, offering us control over the extent of refinement.

Delving into specifics, we initiated with an opening operation, leveraging a 2x2 structuring element to negate diminutive artifacts. This was succeeded by a closing maneuver using a 5x5 structuring element, bridging discontinuities in characters. Subsequent stages involved deploying a 5x11 elliptical structuring element for further gap closure, followed by another opening with a 4x4 structuring element to purge residual noise. This meticulous series culminated in an image inversion, yielding the prepped image crucial for precise character discernment and data extraction.

Our final polishing phase involved a post-processing regimen to obliterate residual noise and discrepancies. We reapplied the opening operation to counter minor artifacts and subsequently employed a closing operation with an enlarged structuring element to bridge any persisting gaps. The terminal inversion rendered the backdrop white and foregrounded the text in black, generating an optimized binary image for the impending OCR. Cumulatively this complex image-processing regimen markedly elevated the fidelity and precision of our OCR outputs.

Figure 2. Process of change with filters and morphological operation.

Figure 3, on its part, shows the results of the original image and the result for the titles in the documents.

Figure 3. Comparison title from original Document and after processing.

Finally, Figure 4 shows the result between the original metadata and the result.

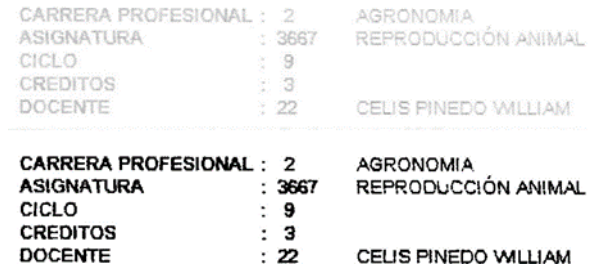


Figure 4. Comparison metadata from original document and after processing.

The quality of the final image in this section of the documents is commend- able, and the OCR process yielded impressive results with a low Character Error Rate (CER) of 2.15% and a Word Error Rate (WER) of 7.05%. These rates demonstrate the high accuracy of the OCR system used and its ability to accurately extract text from the input documents. The low error rates achieved in this study can be attributed to the successful implementation of pre-processing and post-processing techniques, as well as the use of an advanced OCR

tool. These results are promising and indicate the potential of this OCR approach to effectively extract information from noisy scanned PDF documents. We now show the results of the tables segmentation.

Tables

In the realm of document analysis and OCR, image pre-processing is a critical step in enhancing the quality of the input image, which can directly affect OCR accuracy. In this study, we utilized various image processing techniques to improve the readability of the scanned documents, including the removal of lines and other artifacts that can interfere with character recognition. Figure 5 displays the significant improvement in the reconstruction of information from the tables after applying these techniques, as evidenced by the comparison between Figure 5a and Figure 5b. However, it is important to note that the image processing methods employed can sometimes lead to the unintended loss of information, as seen in Figure 5c, where some parts of the image were completely deleted. Nevertheless, the overall benefits of image pre-processing in enhancing OCR accuracy and improving the readability of scanned documents make it a valuable step in the OCR pipeline.

The elimination of lines from the document significantly affected the accuracy of the OCR process. Despite the preprocessing techniques used, the results showed a CER of 10.26% and a WER of 19%. While these values may seem high, it is important to note that OCR accuracy is often impacted by factors such as image quality, font type, and language complexity. Nonetheless, further improvements in the preprocessing techniques and OCR engine could potentially lead to better accuracy in future iterations of the process.

Line extraction

To optimize the input image for OCR, several steps were taken. First, the RGB image was converted to grayscale to reduce the number of color channels, making it easier to work with. This step also removed any color-based noise present in the image. The Gaussian filter was then applied to the grayscale image with a kernel size of 1. The kernel size determines the size of the filter, which smooths the image by averaging the pixel values in the vicinity of each pixel. A small kernel size was used to avoid blurring the edges of the characters. To enhance the thickness of the characters and improve recognition, a dilatation operation was applied to the filtered image using a kernel of size 20x20. The kernel size determines the size of the structuring element used for dilation, which expands the white regions in the image. By using a large kernel, the characters' thickness was increased, making them more visible to the OCR algorithm.

Next, the edges were detected using the canny edge detection algorithm. This algorithm takes two threshold values as inputs, low threshold and high threshold. The values used here were 50 and 150, respectively. These values determine the strength of the edge detection algorithm, with higher values resulting in fewer detected edges. By setting the values to 50 and 150, the algorithm detected only the edges that were most relevant for character recognition, reducing noise in the image.

By applying these pre-processing steps, the image was optimized for OCR, making it easier for the algorithm to detect the characters accurately. The use of appropriate parameters in the Gaussian filter, dilatation operation, and canny edge detection algorithm helped to enhance the contrast and thickness of the characters and reduce noise, improving the accuracy of the OCR system.

Character recognition

Character recognition using Tesseract OCR involves multiple steps. First, the input image is preprocessed to optimize it for OCR. This typically involves converting the image to grayscale, applying filters to remove noise, and enhancing contrast.

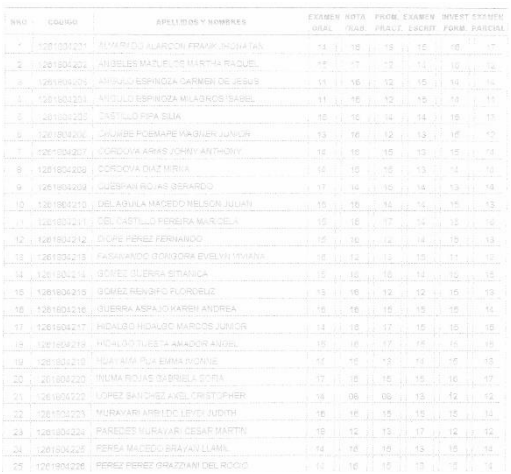
Once the image is preprocessed, Tesseract uses a deep learning-based LSTM (Long Short-Term Memory) architecture to recognize characters in the image. The LSTM network is trained on large datasets of annotated images to learn to recognize a wide variety of character shapes and styles.

During the recognition process, the LSTM architecture performs a sequence of operations on the input image to generate a sequence of probabilities representing the likelihood of each character appearing at each position in the image. These probabilities are then processed using a language model to generate a list of likely candidate words.

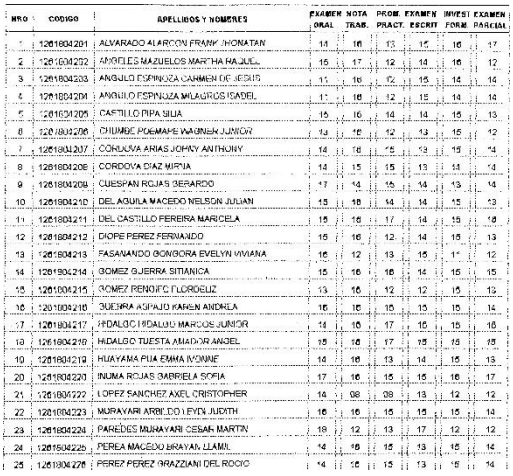
The specific parameters used during character recognition can greatly affect the accuracy of the output. In the provided code snippet, the custom config parameter is used to specify the language and OCR engine mode,

which can affect the recognition accuracy. In this case, the LSTM-based OCR engine is used with LSTM mode 3, which is optimized for word recognition.

Overall, the character recognition process using Tesseract OCR involves multiple complex steps, including image pre-processing, deep learning-based LSTM architecture, and language modeling, to accurately recognize characters and generate text output.

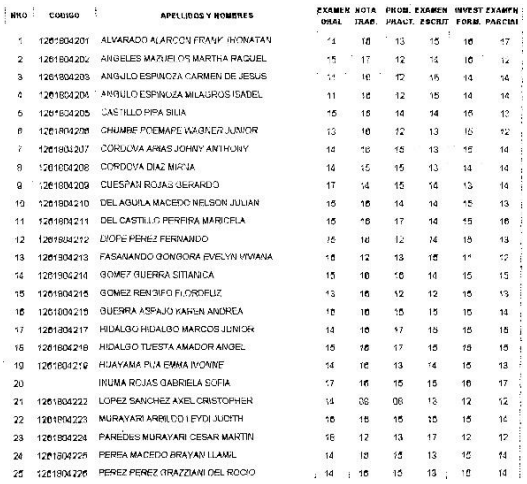


a)



NRO	CODIGO	APELLIDOS Y NOMBRES	EXAMEN ORAL	NOTA TRAB.	PROM. EXAMEN PRACT.	INVEST ESCRIT	EXAMEN FORM.	EXAMEN PARCIAL
1	1261804201	ALVARADO ALARCON FRANK JHONATAN	14	16	13	15	16	17
2	1261804202	ANGELES MAZUELOS MARTHA RAQUEL	15	17	12	14	16	12
3	1261804203	ANGULO ESPINOZA CARMEN DE JESUS	11	16	12	15	14	14
4	1261804204	ANGULO ESPINOZA MILAGROS ISABEL	11	16	12	15	14	14
5	1261804205	CASTILLO PIPA SILIA	15	16	14	14	15	12
6	1261804206	CHUMBE POEMAPE WAGNER JUNIOR	13	16	12	13	15	12
7	1261804207	CORDOVA ARIAS JOHNY ANTHONY	14	16	15	13	15	14
8	1261804208	CORDOVA DIAZ MIRIA	14	15	15	13	14	14
9	1261804209	CUESPAN ROJAS GERARDO	17	14	15	14	13	14
10	1261804210	DEL AGUILA MACEDO NELSON JULIAN	15	16	14	14	15	13
11	1261804211	DEL CASTILLO PEREIRA MARICELA	15	16	17	14	15	16
12	1261804212	DIOPE PEREZ FERNANDO	15	16	12	14	15	13
13	1261804213	FASANANDO GONGORA EVELYN VIVIANA	16	12	13	15	11	12
14	1261804214	GOMEZ GUERRA SITIANICA	15	16	16	14	15	15
15	1261804215	GOMEZ RENGIFO FLORDELIZ	13	16	12	12	15	13
16	1261804216	GUERRA ASPAJO KAREN ANDREA	16	16	15	15	15	14
17	1261804217	HIDALGO HIDALGO MARCOS JUNIOR	14	16	17	15	15	16
18	1261804218	HIDALGO TUESTA AMADOR ANGEL	15	16	17	15	15	15
19	1261804219	HUAYAMA PUA EMMA IVONNE	14	16	13	14	15	13
20	1261804220	INUMA ROJAS GABRIELA SOFIA	17	16	15	15	16	17
21	1261804222	LOPEZ SANCHEZ AXEL CRISTOPHER	14	08	08	13	12	12
22	1261804223	MURAYARI ARBILDO I EYDI JUDITH	15	15	15	15	15	14
23	1261804224	PAREDES MURAYARI CESAR MARTIN	18	12	13	17	12	12
24	1261804225	PEREA MACEDO BRAYAN LLAMIL	14	16	15	13	15	14
25	1261804226	PEREZ PEREZ GRAZZIANI DEL ROCIO	14	16	15	13	15	14

b)



NRO	CODIGO	APELLIDOS Y NOMBRES	EXAMEN ORAL	NOTA TRAB.	PROM. EXAMEN PRACT.	INVEST ESCRIT	EXAMEN FORM.	EXAMEN PARCIAL
1	1261804201	ALVARADO ALARCON FRANK JHONATAN	14	16	13	15	16	17
2	1261804202	ANGELES MAZUELOS MARTHA RAQUEL	15	17	12	14	16	12
3	1261804203	ANGULO ESPINOZA CARMEN DE JESUS	11	16	12	15	14	14
4	1261804204	ANGULO ESPINOZA MILAGROS ISABEL	11	16	12	15	14	14
5	1261804205	CASTILLO PIPA SILIA	15	16	14	14	15	12
6	1261804206	CHUMBE POEMAPE WAGNER JUNIOR	13	16	12	13	15	12
7	1261804207	CORDOVA ARIAS JOHNY ANTHONY	14	16	15	13	15	14
8	1261804208	CORDOVA DIAZ MIRIA	14	15	15	13	14	14
9	1261804209	CUESPAN ROJAS GERARDO	17	14	15	14	13	14
10	1261804210	DEL AGUILA MACEDO NELSON JULIAN	15	16	14	14	15	13
11	1261804211	DEL CASTILLO PEREIRA MARICELA	15	16	17	14	15	16
12	1261804212	DIOPE PEREZ FERNANDO	15	16	12	14	15	13
13	1261804213	FASANANDO GONGORA EVELYN VIVIANA	16	12	13	15	11	12
14	1261804214	GOMEZ GUERRA SITIANICA	15	16	16	14	15	15
15	1261804215	GOMEZ RENGIFO FLORDELIZ	13	16	12	12	15	13
16	1261804216	GUERRA ASPAJO KAREN ANDREA	16	16	15	15	15	14
17	1261804217	HIDALGO HIDALGO MARCOS JUNIOR	14	16	17	15	15	16
18	1261804218	HIDALGO TUESTA AMADOR ANGEL	15	16	17	15	15	15
19	1261804219	HUAYAMA PUA EMMA IVONNE	14	16	13	14	15	13
20		INUMA ROJAS GABRIELA SOFIA	17	16	15	15	16	17
21	1261804222	LOPEZ SANCHEZ AXEL CRISTOPHER	14	08	08	13	12	12
22	1261804223	MURAYARI ARBILDO I EYDI JUDITH	15	15	15	15	15	14
23	1261804224	PAREDES MURAYARI CESAR MARTIN	18	12	13	17	12	12
24	1261804225	PEREA MACEDO BRAYAN LLAMIL	14	16	15	13	15	14
25	1261804226	PEREZ PEREZ GRAZZIANI DEL ROCIO	14	16	15	13	15	14

c)

Figure 5. a) Original Document b) Preprocess image c) Lines erased

Data preparation and exportation

Post-processing is a crucial step that follows OCR, where the extracted text is processed further to correct errors and improve the accuracy of the output. In this step, one potential use of improvement is to correct errors and inconsistencies in the text, such as misspellings, grammatical errors, and formatting issues. Chat-GPT 3, a state-of-the-art language model, has recently been proposed as a tool for post-processing OCR output.

Chat-GPT 3 is a large-scale, transformer-based language model that has achieved remarkable performance across various natural language processing tasks, including language generation, question answering, and language under- standing. The model is trained on a massive amount of text data and can generate coherent and fluent text with high accuracy. In the context of post- processing OCR output, Chat-GPT 3 can be used to correct errors in the extracted text by leveraging its language generation capabilities.

To use Chat-GPT 3 for post-processing OCR output, the extracted text is first passed through the model, which generates a corrected version of the text. This is done by encoding the extracted text into a sequence of tokens using a tokenizer, and then passing the encoded sequence through the model. The model then generates a sequence of tokens that represents the corrected text, which is decoded back into plain text using the same tokenizer. The corrected text can then be used as the final output, improving the accuracy and quality of the OCR output.

To measure the results, two metrics were computed. The Character Error Rate (CER) and Word Error Rate (WER).

$$CER = \frac{S+D+I}{N} \quad (1)$$

Where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of characters. The WER metrics is computed as follow.

$$WER = \frac{S+D+I}{N} \quad (2)$$

Where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words. Table 1, shows the results of each metric.

Table 1. Results. Best value is shown in bold.

Document Id	Proposed approach		Keras OCR	
	CER	WER	CER	WER
1	0.6725	0.2578	0.6258	0.6667
2	0.31624	0.2543	0.5204	0.6968
3	0.2525	0.3121	0.4175	0.6242
4	0.1483	0.2295	0.5599	0.5918
5	1	1	0.6795	0.4979
6	0.7980	0.8653	0.5848	0.6105
7	0.1811	0.3333	0.5401	0.5879
8	1	1	0.4773	0.5416
9	1	1	0.8365	0.9242
10	0.1582	0.25	0.2086	0.3888
11	0.2421	0.2303	0.4067	0.5078

Our study presents significant advances in the field of OCR and NLP, demonstrating a notable improvement in the accuracy of data extraction from former documents. Comparing our results with those of, who achieved an F-score of 0.78 using Pytesseract in the text extraction from Indonesian identification cards, our approach shows an improvement in efficiency and accuracy when processing documents with historical characteristics and of low quality. This is especially relevant considering the unique challenges associated with the digitization n of this type of documents, including variability in the state of conservation and old printing techniques.

The work of Ling, Gao, and Wang (2020) highlights the importance of human-computer interaction and machine learning to improve the efficiency of document processing. Our system incorporates similar principles, optimizing the use of OCR and NLP to reduce processing time and increase accuracy, which is crucial for digital preservation and access to historical documents. Furthermore, our system's ability to adapt to the specificities of former documents significantly improves information retrieval, which reinforces its applicability in digital preservation strategies.

In terms of digital preservation, our study underscores the importance of interoperability with other archival systems. Aligning with the findings of Rosidy, Akhriza, and Husni (2020), who achieved a time efficiency of 94% in extracting information from digital posters, our system proposes a similar solution for historical documents, ensuring that valuable information is accessible and efficiently preserved. Compatibility and ease of integration with other document management platforms are fundamental to achieving a cohesive and effective digital preservation ecosystem.

Based on our results, we agree with , who highlight the utility of OCR in record management to digitize and organize paper-based documents, a principle that our research extends by applying it to historical documents, underlining the need for accurate digitization for their preservation and prolonged access. Simultaneously, Yue, Li, and Hu (2021) demonstrate how intelligent structural analysis improves the precision and efficiency in document data extraction, a strategy that, alongside advances in NLP and OCR in our study, enhances interoperability and integration into digital preservation. Therefore, our work contributes significantly to the field of document processing and digital preservation, highlighting how the adaptability, precision, and interoperability of our system with other archival systems improve the access and preservation of historical documents, aligning with global efforts to protect and facilitate access to our documentary heritage.

Conclusion

The analysis of the metadata and titles revealed a common issue across the dataset where the same error was present in different documents but in varying forms, making it challenging to correct during post-processing. As a result, while the rate of words was high, the rate of character recognition was significantly lower in comparison. The OCR struggled to comprehend certain names and changed them with multiple variations without repeating the elimination or change of characters in a systematic way.

For tables, the OCR model had difficulty in understanding numbers after image recognition, resulting in a high number of errors where numbers were mistakenly converted to letters. Additionally, a problem with both the OCR and the pre-processing stage was the removal of characters from the original restored image, which was found to be significantly higher.

Overall, the application of image processing techniques and NLP proved to be effective in restoring information from noisy scanned PDF documents. However, the study highlighted the limitations of current OCR models, particularly in the recognition of specific characters and numbers. Future research should focus on improving the OCR models to increase the accuracy and efficiency of the data restoration process.

Finally, the methodology proposed in this study has the potential to revolutionize the management of historical educational files, allowing not only for efficient and accurate data retrieval from former records but also their integration into contemporary digital systems. This opens new paths for historical research, academic credential verification, and archive management in institutions with limited resources for digitization. Moreover, the scalability of our technique suggests future applications in broader fields requiring the digitization of historical or low-quality documents, promising to improve access and preservation of valuable documentary heritages globally

References

Ahmad, R., & Rafiq, M. (2023). Global perspective on digital preservation policy: A systematic review. *Journal of Librarianship and Information Science*, *55*(3), 859-867. DOI: https://doi.org/10.1177/09610006221111572

Ahmed, S. F., Alam, M. S. B., Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., ... Gandomi, A. H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, 56, 13521-13617. DOI: https://doi.org/10.1007/s10462-023-10466-8

Barrueco, J. M., & Termens, M. (2022). Digital preservation in institutional repositories: a systematic literature review. *Digital Library Perspectives*, *38*(2), 161-174. DOI: https://doi.org/10.1108/DLP-02-2021-0011

Ghosh, T., Sen, S., Obaidullah, S. M., Santosh, K. C., Roy, K., & Pal, U. (2022). Advances in online handwritten recognition in the last decades. *Computer Science Review*, *46*, 100515. DOI: https://doi.org/10.1016/j.cosrev.2022.100515

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139-172. DOI: https://doi.org/10.1080/23270012.2020.1756939

Kashinath, T., Jain, T., Agrawal, Y., Anand, T., & Singh, S. (2022). End-to-end table structure recognition and extraction in heterogeneous documents. *Applied Soft Computing*, *123*, 108942. DOI: https://doi.org/10.1016/j.asoc.2022.108942

Kim, Y., Choi, S. Y., Park, J., & Kim, J. (2022). Empirical study on human error probability of procedure-extraneous behaviors. *Reliability Engineering and System Safety*, *227*, 108727. DOI: https://doi.org/10.1016/j.ress.2022.108727

Kohli, H., Agarwal, J., & Kumar, M. (2022). An improved method for text detection using Adam optimization algorithm. *Global Transitions Proceedings*, *3*, 230-234. DOI: https://doi.org/10.1016/j.gltp.2022.03.028

Kraus, S., Jones, P., Kailer, N., Weinmann, A., Chaparro-Banegas, N., & Roig-Tierno, N. (2021). Digital Transformation: An Overview of the Current State of the Art of Research. *SAGE Open*, *11*(3), 215824402110475. DOI: https://doi.org/10.1177/21582440211047576

Ling, X., Gao, M., & Wang, D. (2020). Intelligent document processing based on RPA and machine learning. *2020 Chinese Automation Congress (CAC)*, 1349-1353. DOI: https://doi.org/10.1109/CAC51589.2020.9326579

Liu, Y., Wang, Y., & Shi, H. (2023). A Convolutional Recurrent Neural-Network-Based Machine Learning for Scene Text Recognition Application. *Symmetry*, *15*(4), 849. DOI: https://doi.org/10.3390/sym15040849

Lombardi, F., & Marinai, S. (2020). Deep Learning for Historical Document Analysis and Recognition - A Survey. *Journal of Imaging*, *6*(10), 110. DOI: https://doi.org/10.3390/jimaging6100110

Mah, P. M., Skalna, I., & Muzam, J. (2022). Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0. *Applied Sciences*, *12*(18), 9207. DOI: https://doi.org/10.3390/app12189207

Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *IEEE Access*, *8*, 142642-142668. DOI: https://doi.org/10.1109/ACCESS.2020.3012542

Nadkarni, S., & Prügl, R. (2021). Digital transformation: a review, synthesis and opportunities for future research. *Management Review Quarterly*, *71*, 233-341. DOI: https://doi.org/10.1007/S11301-020-00185-7

Patwardhan, N., Marrone, S., & Sansone, C. (2023). Transformers in the Real World: A Survey on NLP Applications. *Information*, *14*(4), 242. DOI: https://doi.org/10.3390/info14040242

Rahali, A., & Akhloufi, M. A. (2023). End-to-End Transformer-Based Models in Textual-Based NLP. *AI*, *4*, 54-110. DOI: https://doi.org/10.3390/ai4010004

Ríos Martínez, N. (2021). Creación de método para el diagnóstico del estado de conservación de documentos de archivos. Caso práctico: fondo documental de la Real Audiencia de Chile. *Intervención*, *2*(24), 92-136. DOI: https://doi.org/10.30763/intervencion.252.v2n24.31.2021

Rosidy, A. S., Akhriza, T. M., & Husni, M. (2020). Combining the NER-OCR methods to improve information retrieval efficiency in the Indonesian posters. *Jurnal Teknologi Dan Sistem Komputer*, *8*(4), 263-269. DOI: https://doi.org/10.14710/jtsiskom.2020.13686

Sulaiman, A., Omar, K., & Nasrudin, M. F. (2019). Degraded Historical Document Binarization: A Review on Issues, Challenges, Techniques, and Future Directions. *Journal of Imaging*, *5*(4), 48. DOI: https://doi.org/10.3390/JIMAGING5040048

Yue, T., Li, Y., & Hu, Z. (2021). DWSA: An Intelligent Document Structural Analysis Model for Information Extraction and Data Mining. *Electronics*, *10*(19), 2443. DOI: https://doi.org/10.3390/electronics10192443

Zanabria-Ortega, M. (2022). Modelo sistémico con enfoque en disciplinas individuales de las organizaciones inteligentes y la eficiencia organizacional. *Revista Científica de Sistemas e Informática*, *2*, e264. DOI: https://doi.org/10.51252/rcsi.v2i1.264

Zeng, G., Zhang, Y., Zhou, Y., Yang, X., Jiang, N., Zhao, G., ... Yin, X.-C. (2023). Beyond OCR + VQA: Towards end-to-end reading and reasoning for robust and accurate textvqa. *Pattern Recognition*, *138*, 109337. DOI: https://doi.org/10.1016/j.patcog.2023.109337