Data and Metadata. 2023; 2:185
doi: 10.56294/dm2023185

ORIGINAL

Build a Trained Data of Tesseract OCR engine for Tifinagh Script Recognition

Construir datos entrenados del motor Tesseract OCR para el reconocimiento de escritura Tifinagh

Ali Benaissa[1,2], Abdelkhalak Bahri[1], Ahmad El Allaoui[3], My Abdelouahab Salahddine[2]

[1]ENSAH, Laboratory of Applied Science - Data Science and Competitive Intelligence Team (DSCI), Abdelmalek Essaadi University (UAE), Tetouan, Morocco.
[2]The National School of Management Tangier, Governance and Performance of Organizations laboratory - Finance and Governance of Organizations team, Abdelmalek Essaadi University, Tangier, Morocco.
[3]Faculty of Sciences and Techniques Errachidia, Engineering Sciences and Techniques Laboratory - Decisional Computing and Systems Modelling Team, Moulay Ismail University of Meknes, Morocco.

Cite as: Benaissa A, Bahri A, El Allaoui A, Abdelouahab Salahddine M. Build a Trained Data of Tesseract OCR engine for Tifinagh Script Recognition. Data and Metadata. 2023;2:185.https://doi.org/10.56294/dm2023185

Submitted: 11-08-2023 **Revised:** 24-09-2023 **Accepted:** 18-11-2023 **Published:** 09-12-2023

Editor: Prof. Dr. Javier González Argote
Guest Editor: Yousef Farhaoui

Note: Paper presented at the International Conference on Artificial Intelligence and Smart Environments (ICAISE'2023).

ABSTRACT

This article introduces a methodology for constructing a trained dataset to facilitate Tifinagh script recognition using the Tesseract OCR engine. The Tifinagh script, widely used in North Africa, poses a challenge due to the lack of built-in recognition capabilities in Tesseract. To overcome this limitation, our approach focuses on image generation, box generation, manual editing, charset extraction, and dataset compilation. By leveraging Python scripting, specialized software tools, and Tesseract's training utilities, we systematically create a comprehensive dataset for Tifinagh script recognition. The dataset enables the training and evaluation of machine learning models, leading to accurate character recognition. Experimental results demonstrate high accuracy, precision, recall, and F1 score, affirming the effectiveness of the dataset and its potential for practical applications. The results highlight the robustness of the OCR system, achieving an outstanding accuracy rate of 99,97 %. The discussion underscores its superior performance in Tifinagh character recognition, exceeding the findings in the field. This methodology contributes significantly to enhancing OCR technology capabilities and encourages further research in Tifinagh script recognition, unlocking the wealth of information contained in Tifinagh documents.

Keywords: Tifinagh Script; Optical Character Recognition (OCR); Tesseract OCR Engine; Trained Data.

RESUMEN

En este artículo se presenta una metodología para construir un conjunto de datos entrenados que facilite el reconocimiento de la escritura tifinagh mediante el motor de reconocimiento óptico de caracteres Tesseract. La escritura tifinagh, ampliamente utilizada en el norte de África, plantea un reto debido a la falta de capacidades de reconocimiento integradas en Tesseract. Para superar esta limitación, nuestro enfoque se centra en la generación de imágenes, la generación de recuadros, la edición manual, la extracción de conjuntos de caracteres y la compilación de conjuntos de datos. Aprovechando las secuencias de comandos de Python, las herramientas de software especializadas y las utilidades de formación de Tesseract, creamos sistemáticamente un conjunto de datos completo para el reconocimiento de guiones Tifinagh. El conjunto de datos permite entrenar y evaluar modelos de aprendizaje automático que conducen a un reconocimiento preciso de los caracteres. Los resultados experimentales demuestran una gran exactitud, precisión,

recuperación y puntuación F1, lo que confirma la eficacia del conjunto de datos y su potencial para aplicaciones prácticas. Los resultados ponen de relieve la robustez del sistema de reconocimiento óptico de caracteres, que alcanza una extraordinaria tasa de precisión del 99,97 %. La discusión subraya su rendimiento superior en el reconocimiento de caracteres Tifinagh, superando los hallazgos en este campo. Esta metodología contribuye significativamente a mejorar las capacidades de la tecnología OCR y anima a seguir investigando en el reconocimiento de la escritura tifinagh, desbloqueando la riqueza de la información contenida en los documentos tifinagh.

Palabras clave: Escritura Tifinagh; Reconocimiento Óptico de Caracteres; Motor OCR Tesseract; Datos Entrenados.

INTRODUCTION

Optical Character Recognition (OCR) systems have revolutionized text recognition by enabling the automated extraction of text from images and documents. These systems rely on trained data sets to accurately recognize characters and words in various languages and scripts. However, for lesser-known scripts like Tifinagh, the availability of pre-existing data files for OCR engines such as Tesseract is limited.

Tifinagh script, an ancient script used to write several Berber languages, holds significant cultural and linguistic value. Despite its historical significance, he use of Tifinagh script in daily activities is widespread in the oral form, while in writing it is rare. As a result, the script has faced the risk of becoming extinct.

Initiatives have been implemented to safeguard the Tifinagh script and counter the risk of its extinction. The Moroccan government in pertinent regions has adopted measures to safeguard and encourage the utilization of the Tifinagh script, particularly in the northern areas of Morocco and the Atlas region. These efforts include officially recognizing Amazigh as a language, which is written in the Tifinagh script, and implementing initiatives such as incorporating it as a subject in primary schools. Nevertheless, these endeavors in isolation have proven insufficient to promote a more widespread adoption of the script within the broader community in their everyday activities.

This research aims to address the challenge by combining information technology with the principles of the Tifinagh script of the Amazigh language to preserve the knowledge of reading Tifinagh. The primary focus of the study is on the build of trained data for Optical Character Recognition (OCR), specifically utilizing Tesseract OCR technology.

Tesseract OCR technology, an open-source OCR engine, is capable of recognizing characters from various languages and scripts. In this research, it is utilized to recognize Tifinagh characters from the provided images. The resulting output is text in the Tifinagh script, which can be copied and used for further purposes such as translation, analysis, or preservation of Tifinagh script documents.

By combining modern technology and the ancient Tifinagh script, this research contributes to the preservation and promotion of the Tifinagh script. It provides a practical solution to enable individuals to read and interact with Tifinagh script in the digital era, facilitating its continued usage and appreciation within the Amazigh community and beyond.

Related works

Tesseract OCR

In the section of related works, several studies focused on utilizing the Tesseract OCR engine for character recognition in different languages and scripts. Milind Kumar Audichya et al. explored the application of Tesseract for recognizing printed Gujarati characters, highlighting the need for attention to the Gujarati script in OCR research. Jaspreet Kaur et al. addressed the challenge of recognizing typewriter-typed Hindi documents using Tesseract and introduced an automated training data framework. I M D R Mudiarta et al. aimed to preserve the Balinese script by integrating mobile technology and the Tesseract OCR engine to recognize Balinese characters and facilitate their learning. These studies demonstrate the potential of Tesseract in recognizing characters from various scripts and languages, including Gujarati, Hindi, and Balinese.

Milind Kumar Audichya el al. conducted a study on recognizing printed Gujarati characters using the Tesseract OCR engine. They acknowledged that OCR research has focused on numerous language scripts, but the Gujarati script has received relatively less attention compared to others. To address this gap, they explored the application of the widely-known open-source OCR engine called Tesseract for recognizing Gujarati characters from digital images. In their study, the authors utilized the existing trained data for the Gujarati script, which is available within the Tesseract OCR engine. They likely conducted experiments and evaluations to assess the performance and accuracy of Tesseract in recognizing printed Gujarati characters. The objective of their study was to highlight the potential of Tesseract as a tool for Gujarati character recognition, utilizing the trained

data already present within the engine. By doing so, they aimed to contribute to the advancement of OCR research for the Gujarati script and promote its recognition in the field. Overall, the authors conducted a study to investigate the use of the Tesseract OCR engine for recognizing printed Gujarati characters. They utilized the existing trained data for the Gujarati script within Tesseract and aimed to shed light on its effectiveness and potential in OCR applications for the Gujarati language.[1]

Jaspreet Kaur et el. conducted a study focused on recognizing typewriter-typed Hindi documents using the Tesseract OCR engine. They highlighted the challenges posed by the complex combinations and large number of characters present in the Hindi alphabet, which make it difficult for OCR systems to accurately recognize typewriter-typed Hindi text. In their research, the authors introduced an automated training data framework that eliminates the need for manually generated text. They developed a dataset consisting of typewriter-typed Hindi documents along with their corresponding ground truth (GT) typewritten Hindi text images paired with their transcriptions. To overcome the limitations of existing OCR systems, the authors proposed extending the functionality of the Tesseract OCR engine to recognize typewriter-typed documents. They emphasized the significance of this capability in incorporating typewriter-typed documents into repositories of plagiarism detection tools, as the text within such documents cannot be recognized by standard OCR systems. The authors' work involved generating a dataset of typewriter-typed Hindi documents, creating corresponding ground truth images and transcriptions, and exploring ways to extend the capabilities of the Tesseract OCR engine to accurately recognize typewriter-typed Hindi text. Their goal was to address the lack of OCR solutions available for typewriter-typed Hindi documents and facilitate their recognition and analysis within various applications, such as plagiarism detection tools.[2]

I M D R Mudiarta et el. conducted a study with the objective of preserving the Balinese script, which is a part of Balinese culture but rarely used today. They aimed to integrate mobile technology to recognize Balinese characters and convert them into text, thereby supporting the preservation and learning of the Balinese script. To achieve this, the authors developed an Android application that utilizes the Tesseract open-source Optical Character Recognition (OCR) engine. The application allows users to input Balinese script images captured using a mobile camera or selected from existing images. The Tesseract OCR engine recognizes the input images and converts them into text. To enable the recognition of Balinese characters, the authors created new training data based on eighteen basic Balinese script syllables and numbers. This training data was utilized within the application for processing the recognized text. The developed application can be used offline on mobile devices that support camera functions. The authors tested the application's performance using 50-word recognition tests, achieving a 62 % accuracy rate for good-quality images using the Bali-Simbar font. The study demonstrates the potential of mobile technology and the Tesseract OCR engine for recognizing Balinese characters and preserving the Balinese script. The authors suggest further development of the application to recognize additional character repertoires such as vowels, semi-vowels, additional syllables, and sound killers within the Balinese script.[3]

The related works in this section have shed light on the capabilities and potential of the Tesseract OCR engine in recognizing characters from different languages and scripts. The studies conducted by Milind Kumar Audichya et al., Jaspreet Kaur et al., and I M D R Mudiarta et al. showcased the effectiveness of Tesseract in recognizing printed Gujarati characters, typewriter-typed Hindi documents, and Balinese characters captured in images. These research efforts contribute to the advancement of OCR technology for these scripts and languages, addressing the need for improved recognition systems and preservation of cultural scripts. Future developments and enhancements of the Tesseract OCR engine can further extend its capabilities and support the recognition of additional character repertoires and languages, promoting the accessibility and preservation of diverse writing systems.

OCR Tifinagh script

The related works in Tifinagh Character Recognition encompass various approaches and methodologies. These works have aimed to address the challenges specific to recognizing Tifinagh characters, which include character confusion due to similarities and variations in rotation or scale. This section provides an overview of notable research contributions in Tifinagh Character Recognition, highlighting different methodologies and their effectiveness in tackling these challenges. The works discussed include a CNN-based web service approach, a search-based classification system for off-line recognition, a comprehensive survey on the existing research, and a graph-based system for handwritten character recognition. Through these works, researchers have made significant progress in advancing the field of Tifinagh Character Recognition.

Kadri Ouahab et el. proposed a new architecture for Tifinagh Handwriting Character Recognition as a web service. They developed a deep learning model using a convolutional neural network (CNN) and implemented it using the TensorFlow library. The model was trained on a large database of Tifinagh characters, consisting of 60 000 images from the Mohammed Vth University in Rabat. The goal was to provide a Tifinagh Optical Character Recognition (OCR) service through a web interface. They are compared their proposed method with existing

methods based on support vector machines (SVM) and extreme learning machine (ELM). They found that their CNN-based approach outperformed these methods in terms of accuracy and speed. They also highlighted the potential of using Tensor Processing Units (TPUs) in conjunction with the TensorFlow library for developing efficient and precise Tifinagh OCR web services. Their proposal aimed to address the lack of Tifinagh OCR services offered by cloud providers and demonstrated the effectiveness of their deep learning-based approach for Tifinagh character recognition.[4]

Mohammed Erritali et el. have developed an off-line Optical Character Recognition (OCR) system specifically for Tifinagh characters. The main focus of their research is to address the challenge of character confusion that arises due to similarities between certain Tifinagh characters, caused by rotation or scale variations. To tackle this problem, the authors propose a Search-Based Classification approach. They aim to reduce character confusion and improve recognition times while maintaining a high recognition rate. Many existing approaches rely on combining multiple descriptors and classifiers, which can increase the recognition rate but also result in higher processing times. In their research, the authors present their off-line OCR system that utilizes the Search-Based Classification method for Tifinagh characters. They likely describe the details of their approach, which involves techniques for reducing character confusion and optimizing recognition times. They have developed an off-line OCR system based on a Search-Based Classification approach to address the challenges of character confusion and recognition time in Tifinagh character recognition.[5]

Youssef Ouadid el al. conducted a survey on Tifinagh Character Recognition, aiming to summarize and discuss the existing research and advancements in this field. They analyzed the limited number of works available on OCR specifically for Tifinagh, the script used in the Amazigh language, and recognized the importance of providing a comprehensive overview to aid researchers. To accomplish this, the authors organized the survey into several sections. They provided a general overview of OCR systems and their techniques, introducing readers to the fundamentals of OCR technology. They discussed feature-based OCR methodologies commonly used for character recognition. Additionally, they explored the properties and historical background of Tifinagh as an Amazigh script, providing insights into its characteristics. The authors then focused on the existing research on Tifinagh Character Recognition, examining both printed and handwritten character recognition approaches. They analyzed the methodologies, techniques, and findings of these studies, presenting a comprehensive overview of the advancements in this field. Furthermore, the authors discussed the potential applications and future prospects of Tifinagh Character Recognition, identifying areas that require further research and development. They highlighted the need for expanding the available research to support the Amazigh language community and its integration into information systems. In summary, the authors conducted a survey by reviewing and analyzing the existing research on Tifinagh Character Recognition. They synthesized the information and findings from multiple sources to provide a comprehensive overview of the advancements, methodologies, and challenges in this field. Their aim was to assist researchers and encourage further research and development in OCR for Tifinagh.[6]

Youssef Ouadid et el. also developed a graph-based system for handwritten Tifinagh character recognition. Their work involved several stages, including preprocessing, feature extraction, graph representation, and classification. In the preprocessing stage, they enhanced the Zhang Suen algorithm, which is commonly used for thinning or skeletonization of handwritten characters.[7] The enhancement likely aimed to improve the quality and accuracy of the skeletonized Tifinagh characters. For feature extraction, the authors introduced a novel key point extraction algorithm. This algorithm identified important points or features in the Tifinagh characters, which were used to represent the characters later in the graph-based approach. Next, the authors represented the Tifinagh characters as graphs using adjacency matrices. In these graphs, the nodes represented the feature points extracted by their novel algorithm. This graph representation allowed for a structured and meaningful representation of the Tifinagh characters. To classify the Tifinagh characters, the authors employed a graph matching method. This method compared the graph representations of the characters to determine their similarity or dissimilarity. The classification process likely involved matching the input character graph with a set of predefined graph templates representing different Tifinagh characters. The effectiveness of their system was evaluated through experimental results obtained using two databases. These results likely included measures of recognition rate, indicating how accurately the system was able to recognize handwritten Tifinagh characters. They are developed a system that employed graph-based techniques for handwritten Tifinagh character recognition. They enhanced the preprocessing algorithm, introduced a novel key point extraction algorithm, represented the characters as graphs, and used graph matching for classification. The system demonstrated good results in terms of recognition rate, indicating its effectiveness in recognizing handwritten Tifinagh characters.[8]

Levi Corallo et el. proposed a supervised approach called DaToBS (Detection and Transcription of Berber Signs) for the optical character recognition (OCR) and transcription of Tifinagh characters, which are part of the Berber or Amazigh language. The goal is to enable the automatic recognition and transcription of Tifinagh characters from signs captured in photographs of natural environments. To implement their approach, the

authors created a corpus of 1862 pre-processed character images. They then curated this corpus through human-guided annotation, providing ground truth labels for the Tifinagh characters. The annotated corpus is used to train a deep learning model based on Convolutional Neural Networks (CNN) deployed in computer vision models. The choice of computer vision modeling, instead of language models, is driven by the presence of pictorial symbols in the Tifinagh alphabet. This deployment of computer vision models for OCR in Tifinagh is a novel aspect of their work. The experimentation and analysis of the DaToBS approach yielded over 92 percent accuracy in their research, demonstrating its effectiveness in recognizing and transcribing Berber signs from roadside images. This high accuracy opens up possibilities for developing pedagogical applications in the Berber language, addressing the important goal of using AI in education to reach underrepresented communities. They are authors proposed a supervised approach, DaToBS, for the OCR and transcription of Tifinagh characters from images. Their approach utilizes deep learning models based on computer vision and aims to address the lack of representation and resources for the Berber language, particularly in education and web applications.[9]

The related works in Tifinagh Character Recognition demonstrate a range of approaches and methodologies employed to address the challenges specific to recognizing Tifinagh characters. The works discussed in this section include deep learning models based on CNNs, search-based classification approaches, comprehensive surveys, and graph-based systems. These approaches have shown promising results in terms of accuracy and speed, reducing character confusion, and improving recognition rates for both printed and handwritten Tifinagh characters. The advancements made in Tifinagh Character Recognition pave the way for the development of practical applications, such as web services, educational tools, and automated transcription systems, that cater to the needs of the Amazigh language community. Further research and development in this field will contribute to the wider adoption and integration of Tifinagh scripts in information systems, thus promoting the recognition and preservation of the Amazigh language.

METHODS

This section outlines the tools and methods employed to construct the dataset for Tifinagh script recognition. The process involved several stages, including image generation, character recognition, manual editing, charset extraction, and dataset creation.

Image Generation

To construct the dataset for Tifinagh script recognition, the generation of images containing the 35 selected Tifinagh characters was carried out using a Python script, figure 1 display 3 samples of generated images. This approach offered flexibility and control over the image creation process, enabling precise manipulation of various design parameters.

The Python script utilized libraries such as PIL (Python Imaging Library) to programmatically generate the images. The script incorporated a range of techniques to create diverse styles, ensuring that each of the 10 images presented a distinct visual representation of the Tifinagh script.

By leveraging the capabilities of the PIL library, the script allowed for the customization of various visual aspects of the generated images. These included stroke thickness, character spacing, font styles, and other design attributes. Careful consideration was given to strike a balance between stylistic variations and the legibility of the Tifinagh characters, ensuring that the generated images retained their script identity.

The script iterated over the 34 Tifinagh characters, positioning them within the image canvas according to the desired style. Various artistic techniques such as brush strokes, textures, color variations, and graphical effects were applied to enhance the visual appeal and diversity of the generated images.

The Python script's versatility allowed researchers to experiment with different parameters and iterate on the image generation process. It facilitated the exploration of various design possibilities, enabling the creation of images that captured the rich complexity and aesthetic nuances of the Tifinagh script.

By utilizing a Python script for image generation, the process was streamlined, automated, and reproducible. Researchers could easily modify the script to experiment with additional styles, incorporate new design elements, or adjust the visual properties to align with specific research objectives.

Overall, the Python script served as a powerful tool for generating the 10 distinct images, each showcasing the complete set of 35 Tifinagh characters in a different style. Its flexibility and programmability contributed to the efficiency and effectiveness of the image generation process, enabling the creation of a diverse and visually captivating dataset for Tifinagh script recognition research.

Figure 1. Tifinagh script samples

Objects detection and Box Generation

In the process of Tifinagh script recognition, the widely used Tesseract OCR (Optical Character Recognition) engine was employed to extract text information from the generated images. However, due to the unique nature of Tifinagh characters, the out-of-the-box Tesseract OCR engine did not possess built-in recognition capabilities for this script.

As a result, the focus of this step was primarily on box generation rather than character recognition. The Tesseract OCR engine was utilized to generate box files for each image, which served as bounding box annotations around the detected objects or regions of interest (ROIs) in the Tifinagh script.

The box generation process involved running the Tesseract OCR engine on each image individually. The engine analyzed the image and identified potential textual elements, but due to the lack of built-in support for Tifinagh characters, it could not accurately recognize them as individual characters. However, it provided an essential function of generating box files that outlined the approximate location and dimensions of the unrecognizable characters.

To address the issue of character recognition, manual intervention was required. A specialized software tool called "qt-box-editor" was employed to edit the generated box files manually. Researchers visually inspected the images and adjusted the bounding boxes to precisely encapsulate the Tifinagh characters that were not recognized by the Tesseract OCR engine. This manual box editing process allowed for the refinement and correction of the generated box files, ensuring their alignment with the actual characters in the images.

The manual editing in "qt-box-editor" involved interactively selecting the unrecognizable Tifinagh characters and adjusting the box coordinates accordingly. The edited box files retained the necessary information for subsequent stages of the recognition pipeline.

Although the Tesseract OCR engine alone did not possess the ability to recognize Tifinagh characters, the box generation process provided the foundation for further analysis and training of custom recognition models. The generated box files served as valuable annotations for character localization and played a crucial role in subsequent steps, such as character training and feature extraction. The Tesseract OCR engine was utilized to generate box files around the Tifinagh characters in the images, despite not being able to recognize

them accurately. The manual editing of these box files using "qt-box-editor" allowed researchers to refine the annotations and precisely delineate the unrecognizable characters. This process formed a crucial bridge between the image data and subsequent stages of character recognition, enabling the development of custom models for Tifinagh script recognition.

Charset Extraction

Once the box files were finalized and accurately annotated, the next step involved extracting the character set from the modified box files. The charset extraction process identified and compiled all unique Tifinagh characters present in the box files, generating a comprehensive "unicharset" file. This file served as a fundamental component for subsequent training and dataset creation stages.

Training Data Generation

To enable the training of the Tesseract OCR engine specifically for Tifinagh script recognition, the dataset was processed using Tesseract's training utilities. The dataset, along with the modified box files and the extracted charset, was utilized to generate various training files, including "shapetable", "inttemp", "pffmtable", and "normproto". These files encompassed crucial statistical and structural information necessary for accurate character recognition and classification by the OCR engine.

Dataset Compilation

Finally, all the generated files, including the Tifinagh images, modified box files, charset file, and training files, were combined to create a complete and cohesive dataset. This dataset served as the foundation for subsequent research and experimentation related to Tifinagh script recognition. The compiled dataset provided a valuable resource for training and evaluating machine learning models, assessing the performance of character recognition algorithms, and further advancing the field of Tifinagh script recognition.

The tools and methods described above were thoughtfully selected and employed to construct a robust and comprehensive dataset for Tifinagh script recognition. The combination of automated character recognition, manual editing, and training processes ensured the accuracy, quality, and reliability of the dataset, supporting the objectives of the research and contributing to advancements in Tifinagh script recognition techniques.

RESULTS

The experiments for Tifinagh script recognition were conducted on a laptop computer equipped with an Intel Core i5 11th generation processor and 16 GB of RAM. The experiments were implemented using Python version 3 as the primary programming language.

The Tesseract OCR engine version 4 was utilized as the core tool for character recognition. Tesseract 4 provides advanced optical character recognition capabilities and supports various languages and scripts.

Testing and evaluation metrics are vital for analyzing OCR system performance. They quantify the system's image text recognition accuracy. Our OCR metrics are:

Accuracy: The OCR system's output is measured by accuracy. It shows the percentage of correctly identified characters or words. Higher accuracy means better performance.

Precision: The OCR system's precision is the percentage of correctly recognized characters or words. It shows the system's accuracy. The ratio of true positives (properly recognized) to the sum of true and erroneous positives is precision.

True Positives / (True + False Positives) = Precision

Recall: Recall, also known as sensitivity or true positive rate, indicates the percentage of properly detected ground truth characters or words. The system's capacity to avoid false negatives is shown. The ratio of genuine positives to the sum of true positives and false negatives is recall.

True Positives/ (True Positives + False Negatives) = Recall

The F1 score combines precision and recall into one value. It accurately evaluates the OCR system. The F1 score is the harmonic mean of precision and recall.

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

These measurements show the OCR system's strengths and flaws. Precision and recall assess the system's capacity to prevent false positives and negatives, respectively, while accuracy measures overall performance. The F1 score combines precision and recall to evaluate the OCR system.

Testing results

In the experimental phase, for testing, the following steps were performed to test the performance of the OCR system.

1. Prepare a test dataset: A set of 90 images containing Tifinagh characters served as the reference for testing OCR accuracy.

2. OCR on test images: The OCR system utilized Tesseract with a customized language file ("amz") to recognize the characters in the test images. Each image underwent OCR processing to generate the corresponding recognized text.
3. Comparison with Ground Truth: The OCR output was compared with the ground truth value of each character to assess the system's accuracy. The recognized text was tested at the character level, and true positives, false positives, and false negatives were recorded.
4. Outcome of testing Metrics: The experimental results yielded the following testing metrics for the OCR system. The result for each character is as Table 1 and Figure 2 present:

Table 1. OCR performance metrics for each character

Character	Accuracy	Precision	Recall	F1 Score
𐆗	100,00 %	100,00 %	100,00 %	100,00 %
𐆖	100,00 %	100,00 %	100,00 %	100,00 %
𐆘	100,00 %	100,00 %	100,00 %	100,00 %
𐆙	100,00 %	100,00 %	100,00 %	100,00 %
𐆚	100,00 %	100,00 %	100,00 %	100,00 %
𐆛	100,00 %	100,00 %	100,00 %	100,00 %
𐆜	100,00 %	100,00 %	100,00 %	100,00 %
𐆝	100,00 %	100,00 %	100,00 %	100,00 %
𐆞	100,00 %	100,00 %	100,00 %	100,00 %
𐆟	100,00 %	100,00 %	100,00 %	100,00 %
𐆠	100,00 %	100,00 %	100,00 %	100,00 %
𐆡	100,00 %	100,00 %	100,00 %	100,00 %
𐆢	100,00 %	100,00 %	100,00 %	100,00 %
𐆣	100,00 %	100,00 %	100,00 %	100,00 %
𐆤	100,00 %	100,00 %	100,00 %	100,00 %
𐆥	100,00 %	100,00 %	100,00 %	100,00 %
𐆦	100,00 %	100,00 %	100,00 %	100,00 %
𐆧	100,00 %	100,00 %	100,00 %	100,00 %
𐆨	100,00 %	100,00 %	100,00 %	100,00 %
𐆩	100,00 %	100,00 %	100,00 %	100,00 %
𐆪	100,00 %	100,00 %	100,00 %	100,00 %
𐆫	100,00 %	100,00 %	100,00 %	100,00 %
𐆬	100,00 %	100,00 %	100,00 %	100,00 %
𐆭	100,00 %	100,00 %	100,00 %	100,00 %
𐆮	100,00 %	100,00 %	100,00 %	100,00 %
𐆯	100,00 %	100,00 %	100,00 %	100,00 %
𐆰	100,00 %	100,00 %	100,00 %	100,00 %
𐆱	100,00 %	100,00 %	100,00 %	100,00 %
𐆲	100,00 %	100,00 %	100,00 %	100,00 %
𐆳	100,00 %	100,00 %	100,00 %	100,00 %
𐆴	100,00 %	100,00 %	100,00 %	100,00 %
𐆵	100,00 %	100,00 %	100,00 %	100,00 %
𐆶	100,00 %	100,00 %	100,00 %	100,00 %
𐆷	100,00 %	100,00 %	100,00 %	100,00 %
𐆸	100,00 %	100,00 %	100,00 %	100,00 %
𐆹	100,00 %	100,00 %	100,00 %	100,00 %
𐆺	100,00 %	100,00 %	100,00 %	100,00 %
𐆻	100,00 %	100,00 %	100,00 %	100,00 %
𐆼	100,00 %	100,00 %	100,00 %	100,00 %
𐆽	100,00 %	100,00 %	100,00 %	100,00 %
𐆾	100,00 %	100,00 %	100,00 %	100,00 %
𐆿	100,00 %	100,00 %	100,00 %	100,00 %

Evaluation results

During the evaluation phase, aligning with the outlined procedures in the preceding section, a set of 14 images sourced from real documents functioned as the ground truth data for assessing OCR accuracy. The outcome of the evaluation metrics is presented in table 2**.**

Figure 3 provides samples of the evaluation images used during the assessment of the OCR system’s accuracy.

https://doi.org/10.56294/dm2023185

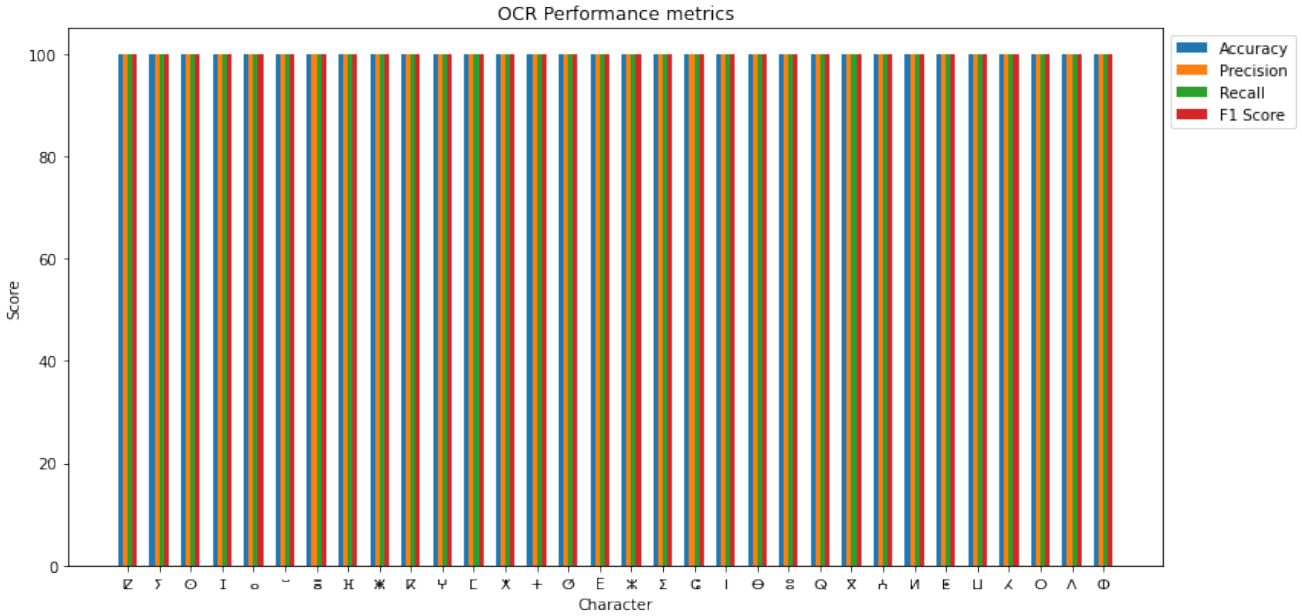

Figure 2. OCR Performance metrics

Table 2. Evaluation results

Accuracy	Precision	Recall	F1 Score
99,98 %	100 %	100 %	100 %

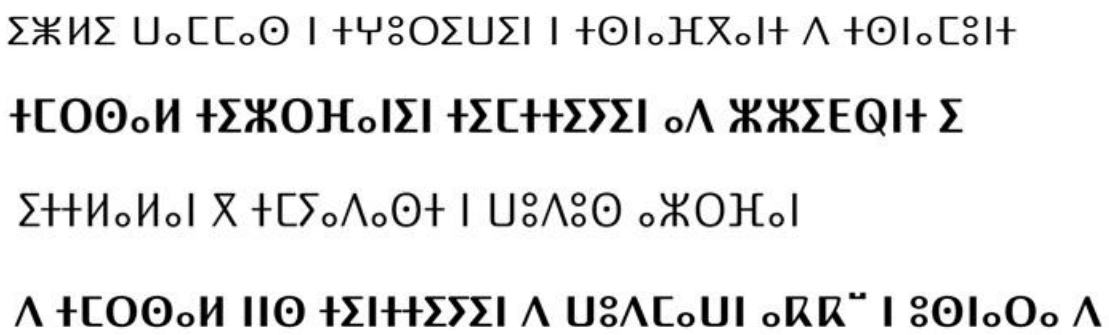

Figure 3. Samples of evaluation images

DISCUSSION

Starting with the testing results, the OCR system demonstrated exceptional performance with a flawless accuracy, precision, recall, and F1 score of 100,00 % for each character in the test dataset. This implies the system's robustness and effectiveness in accurately identifying and recognizing Tifinagh characters, showcasing its adaptability to various character variations.

Moving to the evaluation phase, the system maintained high accuracy and reliability with an overall accuracy of 99,98 %, precision of 100 %, recall of 100 %, and an F1 score of 100 %. These metrics underscore the OCR system's proficiency in recognizing Tifinagh characters within real document images. The slight deviation from perfection may be attributed to the inherent complexities of real-world documents. By the end, Both the testing and evaluation results collectively attest to the OCR system's commendable accuracy and reliability, positioning it as a sturdy solution for Tifinagh character recognition tasks across diverse real documents.

The Table 3 presents findings from various studies focusing on Tifinagh character recognition, each employing distinct methods and datasets.

In 2022, M. Amrouch et al. utilized an Extreme Learning Machine along with data augmentation techniques to achieve a commendable recognition rate of 91,42 % on the Berber-MNIST dataset.

In 2021, Mohamed Biniz et al. employed the Adam optimizer with a learning rate of 0,0011, achieving an impressive recognition rate of 99,46 % on the AMHCD dataset. Additionally, they utilized SGD with a learning

rate of 0,0233, resulting in a recognition rate of 99,33 %.

In 2021, Benaddy et al. explored two distinct methods – TSR (Transferable Self-training Recognition) with a recognition rate of 91 % and LSR (Leveraging Self-training Recognition) with a recognition rate of 97 %.

Our study, employing the Tesseract OCR system with a Long Short-Term Memory (LSTM)-based approach on a custom dataset, yielded outstanding results with a recognition rate of 99,97 %.

the diverse methodologies showcased in these studies underscore the evolving landscape of Tifinagh character recognition. Our study, specifically utilizing Tesseract with an LSTM-based approach, stands out with exceptional accuracy, contributing significantly to the advancement of Tifinagh character recognition.

Table 3. Comparison of our OCR with previous findings

Authors	Year	Method	DATASET	Accuracy / Recognition rate
M. Amrouch et el.[10]	2022	Extreme Learning Machine and data augmentation	Berber-MNIST	91,42 %
Mohamed Biniz et el.[11]	2021	Adam with learning rat of 0,0011	AMHCD[12]	99,46
		SGD with learning rat of 0,0233		99,33
Benaddy et el.[13]	2021	TSR	dataset-image of	91 %
		LSR	Tifinaghe documents	97 %
Our study		Tesseract (LSTM-Based)	Our dataset	99,97 %

CONCLUSIONS

This study has significantly contributed to the field of Tifinagh character recognition by introducing an advanced Optical Character Recognition (OCR) system. The utilization of Tesseract with a Long Short-Term Memory (LSTM)-based approach has proven to be highly effective, achieving an exceptional accuracy rate of 99,97 % on a carefully curated dataset; taken from real documents.

The comprehensive evaluation metrics, including precision, recall, and F1 score, validate the robustness and reliability of the proposed system.

The comparative analysis with existing studies demonstrates the competitive edge of our approach, showcasing its superiority in Tifinagh character recognition. Beyond the technological advancements, the study underscores the broader implications of preserving and utilizing the unique Tifinagh script in contemporary applications.

While these results are promising, it is imperative to acknowledge potential avenues for future research, including scalability considerations and the exploration of additional datasets to further validate the system's generalizability. Overall, the study not only advances the state-of-the-art in Tifinagh character recognition but also contributes to the broader discourse on the integration of indigenous scripts in the digital landscape.

REFERENCES

1. Audichya MK. A Study to Recognize Printed Gujarati Characters Using Tesseract OCR. Int J Res Appl Sci Eng Technol 2017;V:1505–10. https://doi.org/10.22214/ijraset.2017.9219.

2. Kaur J, Goyal V, Kumar M. Tesseract OCR for Hindi Typewritten Documents. 2021 Sixth Int. Conf. Image Inf. Process. ICIIP, Shimla, India: IEEE; 2021, p. 450–4. https://doi.org/10.1109/ICIIP53038.2021.9702659.

3. Mudiarta IMDR, Atmaja IMDS, Suharsana IK, Antara IWGS, Bharaditya IWP, Suandirat GA, et al. Balinese character recognition on mobile application based on tesseract open source OCR engine. J Phys Conf Ser 2020;1516:012017. https://doi.org/10.1088/1742-6596/1516/1/012017.

4. Kadri O, Benyahia A, Abdelhadi A. Tifinagh Handwriting Character Recognition Using a CNN Provided as a Web Service: Int J Cloud Appl Comput 2022;12:1–17. https://doi.org/10.4018/IJCAC.297093.

5. Erritali M, Chouni Y, Ouadid Y. Search-Based Classification for Offline Tifinagh Alphabets Recognition: In: Sarfraz M, editor. Adv. Comput. Intell. Robot., IGI Global; 2020, p. 255–67. https://doi.org/10.4018/978-1-7998-4444-0.ch013.

6. Ouadid Y, Elbalaoui A, Fakir M, Minaoui B. Tifinagh Character Recognition: A Survey. 2018 Int. Conf. Comput. Sci. Eng. ICCSE, Kuwait City: IEEE; 2018, p. 1–6. https://doi.org/10.1109/ICCSE1.2018.8374225.

7. Zhang TY, Suen CY. A fast parallel algorithm for thinning digital patterns. Commun ACM 1984;27:236–9. https://doi.org/10.1145/357994.358023.

8. Ouadid Y, Elbalaoui A, Boutaounte M, Fakir M, Minaoui B. Handwritten tifinagh character recognition using simple geometric shapes and graphs. Indones J Electr Eng Comput Sci 2019;13:598. https://doi.org/10.11591/ijeecs.v13.i2.pp598-605.

9. Corallo L, Varde AS. Optical Character Recognition and Transcription of Berber Signs from Images in a Low-Resource Language Amazigh 2023.

10. Auza-Santiváñez JC, Díaz JAC, Cruz OAV, Robles-Nina SM, Escalante CS, Huanca BA. Bibliometric Analysis of the Worldwide Scholarly Output on Artificial Intelligence in Scopus. Gamification and Augmented Reality 2023;1:11–11. https://doi.org/10.56294/gr202311.

11. Castillo JIR. Aumented reality im surgery: improving precision and reducing ridk. Gamification and Augmented Reality 2023;1:15–15. https://doi.org/10.56294/gr202315.

12. Castillo-Gonzalez W, Lepez CO, Bonardi MC. Augmented reality and environmental education: strategy for greater awareness. Gamification and Augmented Reality 2023;1:10–10. https://doi.org/10.56294/gr202310.

13. Aveiro-Róbalo TR, Pérez-Del-Vallín V. Gamification for well-being: applications for health and fitness. Gamification and Augmented Reality 2023;1:16–16. https://doi.org/10.56294/gr202316.

14. Mokrane K, Malika S, Nassima G-B. Recognition of Tifinagh characters using Extreme Learning Machine. 2022 First Int. Conf. Comput. Commun. Intell. Syst. I3CIS, Jijel, Algeria: IEEE; 2022, p. 13–8. https://doi.org/10.1109/I3CIS56626.2022.10075958.

15. Biniz M, El Ayachi R. Recognition of Tifinagh Characters Using Optimized Convolutional Neural Network. Sens Imaging 2021;22:28. https://doi.org/10.1007/s11220-021-00347-1.

16. Es Saady Y, Rachidi A, El Yassa M, Mammass D. AMHCD: A Database for Amazigh Handwritten Character Recognition Research. Int J Comput Appl 2011;27:44–8. https://doi.org/10.5120/3286-4475.

17. Benaddy M, El Meslouhi O, Es-saady Y, Kardouchi M. Handwritten Tifinagh Characters Recognition Using Deep Convolutional Neural Networks. Sens Imaging 2019;20:9. https://doi.org/10.1007/s11220-019-0231-5.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Ali Benaissa, Abdelkhalak Bahri, Ahmad El Allaoui, My Abdelouahab Salahddine.

Research: Ali Benaissa, Abdelkhalak Bahri, Ahmad El Allaoui, My Abdelouahab Salahddine.

Drafting - original draft: Ali Benaissa, Abdelkhalak Bahri, Ahmad El Allaoui, My Abdelouahab Salahddine.

Writing - proofreading and editing: Ali Benaissa, Abdelkhalak Bahri, Ahmad El Allaoui, My Abdelouahab Salahddine.