IndicOCR: A Pipeline for Recognizing Printed Documents for Indian Languages

Krishna Tulsyan
krishna.tulsyan@research.iiit.ac.in
International Institute of Information Technology
Hyderabad, Telangana, India

Tessy Flemin
tessy.flemin@research.iiit.ac.in
International Institute of Information Technology
Hyderabad, Telangana, India

Ajoy Mondal
ajoy.mondal@iiit.ac.in
International Institute of Information Technology
Hyderabad, Telangana, India

C V Jawahar
jawahar@iiit.ac.in
International Institute of Information Technology
Hyderabad, Telangana, India

ABSTRACT

India's linguistic diversity is a testament to the cultural richness of the subcontinent, with 22 officially recognized languages, each possessing its unique script and identity. In this context, the need for efficient Optical Character Recognition (OCR) tools tailored to the complexities of these languages is paramount. **IndicOCR** is introduced as an innovative OCR pipeline designed to address this challenge. This paper aims to shed light on the capabilities of **IndicOCR**, underlining its role in bridging the gap between India's linguistic heritage and the ever-expanding digital world. Finally, **IndicOCR**[1] is a powerful tool for inclusivity, preservation, and progress in a linguistically diverse society.

CCS CONCEPTS

• **Software and its engineering** → Software version control; **Software infrastructure**; *Layered systems*; *API languages*; Object oriented frameworks.

KEYWORDS

Optical Character Recognition (OCR), Indian Script/Language, Printed Text, Unicode Character.

ACM Reference Format:
Krishna Tulsyan, Tessy Flemin, Ajoy Mondal, and C V Jawahar. 2024. IndicOCR: A Pipeline for Recognizing Printed Documents for Indian Languages. In *7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD) (CODS-COMAD 2024), January 04–07, 2024, Bangalore, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3632410.3632502

1 INTRODUCTION

India, a land of diversity and cultural richness, boasts a linguistic tapestry that is as vibrant as it is complex. The country's linguistic landscape presents a unique challenge and opportunity with 22 officially recognized languages, each with script and dialectical variations. The need for technology that can bridge the gap between this rich linguistic heritage and the digital world has never been more pronounced. In response to this challenge, we present IndicOCR, a demo pipeline designed to recognize printed documents in 22 Indian languages.

The development of Optical Character Recognition (OCR) technology has revolutionized the digitization of text-based content worldwide. However, this transformative technology has only sometimes been tailored to the specific needs of diverse linguistic regions like India. The intricate scripts, various character sets, and varying text orientations across Indian languages have posed significant obstacles to accurately and efficiently recognizing printed documents.

This paper presents the IndicOCR pipeline, a comprehensive solution showcasing our commitment to inclusivity and accessibility in the digital realm. The core components of the IndicOCR pipeline include (i) **Script/Language Identification:** the ability to detect the script of a given text automatically, a crucial step in accurately recognizing Indian languages, (ii) **Word Segmentation:** accurate segmentation of words, particularly in scripts with complex ligatures and conjuncts, is essential for IndicOCR, (iii) **Word Recognition:** the heart of IndicOCR, our word recognizer module utilizes deep neural networks to recognize words in the document, (iv) **Post OCR Error Correction:** tailored post-processing technique and language-specific support ensure high accuracy across the 22 languages, and (v) **Application Interface:** IndicOCR is designed with usability in mind, offering a user-friendly interface that enables individuals and organizations to integrate OCR capabilities into their workflows easily. Moreover, we will provide practical demonstrations of IndicOCR's capabilities, highlighting its adaptability to various printed documents, including books, newspapers, magazines, and official records.

As we embark on this journey through the IndicOCR pipeline, we invite you to explore the convergence of technology and culture, witness the democratization of information access, and envision a future where the digital realm embraces linguistic pluralism. IndicOCR represents more than just a technical achievement; it is a testament to our commitment to fostering an inclusive and digitally empowered India, where every language has a place in the digital world.

[1]Demo site: https://ilocr.iiit.ac.in/indicocr/

CODS-COMAD 2024, January 04–07, 2024, Bangalore, India

ACM ISBN 979-8-4007-1634-8/24/01.
https://doi.org/10.1145/3632410.3632502

2 INDICOCR

2.1 Core Component

The core components of the IndicOCR include (i) **Script/Language Identification**, (ii) **Pre-processing**, (iii) **Word Segmentation**, (iv) **Word Recognition**, (v) **Post OCR Error Correction**. These core components collectively enable the recognition of printed documents in the diverse languages of India, making it a powerful tool for digitization, accessibility, and preservation of linguistic and cultural heritage in the digital age.

Language/Script Identification: This component identifies the language/script of input documents. Selecting the appropriate recognition model and language-specific post-error correction for accurate recognition is crucial. We train a deep classification network [2] with word-level images of 22 Indian languages for this task.

Pre-processing: This module is essential for enhancing the quality of the input document image. It typically involves noise reduction, contrast enhancement, and image binarization to enhance input document images for better recognition. For this purpose, we use Conditional Generative Adversarial Network (GAN)-based document enhancement technique [4].

Word Segmentation: In this phase, the input document image is divided into individual words. The system employs a fine-tuned docTR [3] model to achieve this goal. By leveraging word bounding boxes, the IndicOCR system can accurately delineate the boundaries of words in the document image. This module ensures that each word is processed individually, improving the accuracy of the recognition module.

Word Recognition: This module is the heart of our OCR pipeline. We leverage a Convolutional Recurrent Neural Network (CRNN) developed by Gongidi *et al.* [1], which consists of four modules: Transformation Network (TN), Feature Extractor (FE), Sequence Modeling (SM), and Predictive Modeling (PM). The transformation network comprises six plain convolutional layers with varying numbers of channels (16, 32, 64, 128, 128, and 128). Each layer has a filter size of 3, a stride of 1, and padding of 1. A 2x2 max-pooling layer with a stride of 2 follows each convolutional layer. We use the ResNet architecture for the feature extractor, which is known for its effectiveness in feature extraction. The sequence modeling module has a 2-layer Bidirectional LSTM (BLSTM) architecture, each with 256 hidden neurons. Lastly, the predictive modeling module utilizes Connectionist Temporal Classification (CTC) to decode and recognize words.

Post OCR Error Correction: After the recognition of words, post-processing techniques are applied to correct the errors and enhance the accuracy of the recognized text. This step involves language-specific spell-checker correction tools.

2.2 Application Interface

IndicOCR is an online web application designed for exceptional usability, empowered by advanced technology. It utilizes **FastAPI** for efficient back-end processing and relies on **ReactJS** for its user-friendly front-end interface. Here is a deeper exploration of what this entails:

User Interface: IndicOCR adopts a user-centric approach by providing a lightning-fast and highly responsive user interface. Powered by ReactJS, it ensures that users experience a smooth and efficient interaction with the application. Whether processing a single document or managing a large volume of images, IndicOCR remains fast and responsive. Its intuitive user interface further enhances the user experience, making document management and OCR tasks seamless and efficient. Additionally, the system is scalable to effortlessly handle the processing of thousands of images, ensuring it meets the demands of even the most extensive document digitization projects.

Streamlined Workflow Integration: FastAPI is a powerful tool that enables seamless integration of OCR capabilities into users' workflows. It is equally useful for individuals managing personal documents and organizations automating data extraction from archives. This technical foundation makes data processing and incorporation into existing workflows a breeze, saving users significant time and effort in document digitization and data extraction tasks.

Customization Options: IndicOCR provides a range of customization options to meet the diverse needs of its users. With granular control over critical OCR parameters and settings, users can personalize the OCR process to match their specific requirements.

- **Language:** Users can select the language or languages in which the OCR engine should recognize text. This flexibility ensures that IndicOCR accurately processes documents in various languages and scripts.
- **Modality:** IndicOCR allows users to specify the document modality, whether it is a printed document, handwritten text, or a combination of both. This modality customization enhances the OCR engine's ability to handle different types of content effectively.
- **Pre-processing:** Users can apply pre-processing filters to enhance document quality before OCR. These filters can include noise reduction, contrast adjustment, and image enhancement, ensuring optimal OCR results.
- **Region Selection:** Users can manually specify a rectangular text region on the image being processed. This level of precision allows users to focus OCR on specific areas of interest within the document, excluding irrelevant content and ensuring accuracy where it matters most.
- **Segmentation Model:** Users can choose from different segmentation models to define how the OCR engine should divide text and elements within a document. This customization is particularly valuable for complex documents with multiple columns, headers, and footers.
- **Recognition Model:** IndicOCR offers a selection of recognition models, each optimized for specific text types or styles. Users can fine-tune this setting to achieve the highest accuracy for their particular documents.
- **Post-processing:** After OCR, users can apply post-processing rules to refine the extracted text further. It includes spell-checking, de-duplication, and formatting adjustments to ensure the output meets their requirements.

These comprehensive customization options empower users to adapt IndicOCR's OCR capabilities to their unique document processing needs. Users can achieve highly accurate and tailored OCR results for their specific use cases by selecting the correct language, optimizing pre-processing filters, or fine-tuning segmentation and recognition models.

Versatility: IndicOCR showcases exceptional versatility by accommodating a multitude of input formats. It effortlessly processes images of any type, making it an ideal solution for digitizing text from photographs, scanned documents, or graphical content. Additionally, IndicOCR efficiently manages PDF files, automatically converting them into individual images before applying OCR. This feature simplifies text extraction from PDF documents and ensures users can seamlessly work with different document sources.

Moreover, the application also offers flexible output options. Users can choose from various output formats based on their needs, whether you require plain text for easy readability, structured data in JSON for integration with other applications, or highly structured content in XML. This extensive range of output formats enhances the usability of the extracted data across a broad spectrum of applications.

This versatility extends the utility of IndicOCR across diverse industries and use cases. Whether you are in research, data analysis, document archiving, or content management, IndicOCR's ability to handle different input sources and provide data in various formats ensures that it seamlessly integrates into your workflow, facilitating efficient data extraction and processing.

Seamless Updates: IndicOCR embraces a transparent and dynamic update approach, leveraging version control using Git. It allows users to track changes, access historical versions, and stay informed about the evolution of the application.

Its segmentation and recognition models are frequently improved, enhancing accuracy and expanding language support. The application benefits from a vibrant open-source community where developers and users collaborate to refine and innovate its features and capabilities.

With this open-source foundation, IndicOCR will respond to user feedback and evolve rapidly to incorporate the latest developments in the OCR field. It ensures that users will have access to the most up-to-date advancements and can actively contribute to the application's growth and adaptability in the future.

Documentation and Support: IndicOCR provides comprehensive documentation that guides users through its features and functionalities. This documentation is a valuable resource for users to learn how to make the most of the application, troubleshoot common issues, and explore advanced capabilities. Accessible and detailed documentation ensures that users can quickly get started and maximize their productivity with IndicOCR.

In conclusion, IndicOCR is a technologically advanced OCR solution that seamlessly combines technical prowess in the backend with a user-centric and intuitive front-end experience. This approach allows users to harness the full potential of OCR capabilities while benefiting from a streamlined and user-friendly interface.

Miscellaneous Features. IndicOCR goes beyond its core functions with additional small but essential features. These include robust authentication to ensure secure access, advanced permission handling for controlling document access, extensive logging for tracking processing history, and the convenience of bulk upload and download for efficient document management. These extras enhance the overall user experience and make IndicOCR a comprehensive solution for document-related tasks.

3 INDICOCR DEMONSTRATION

Fig.1 shows the pipeline for extracting textual content from a document image through IndicOCR. An online version of the software is available at https://ilocr.iiit.ac.in/indicocr/. A demo video can be found at https://youtube.com/watch?v=-wiyEoeQElI&si=bYAdJJ14oYjuy48C.

4 POTENTIAL APPLICATIONS

The Indic languages, also known as the Indo-Aryan languages, hold significant cultural, historical and linguistic importance. It showcases rich cultural heritage diversity and enables widespread usage and connectivity. The linguistic diversity in these languages has, by and large, attracted many linguists, scholars, and researchers. Many indigenous practices and traditional knowledge are documented in the Indic languages. Preserving and promoting these languages is crucial for safeguarding this invaluable heritage.

IndicOCR is essential for promoting digital inclusion, preserving cultural heritage, improving efficiency in various sectors, and ensuring that Indic-speaking populations have equitable access to information and services in their native languages globally.

Some of the potential applications of IndicOCR include:

Digitization of Historical Texts: Many historians/archivists make additional efforts to preserve historical documents as part of their academic or personal pursuits. They are dedicated to collecting, cataloging, preserving, and providing access to historical documents, manuscripts, photographs, and other valuable materials. To ensure that their work and efforts reach a widespread audience, digitization of these contents is extremely important. IndicOCR plays a significant role by helping digitize the contents, even for transliterated content such as 'Sanskrit has written in Telugu/Hindi'.

IndicOCR can convert ancient documents written in Indic scripts into digital text. The text becomes searchable. Researchers, scholars, and enthusiasts can easily search for specific keywords or phrases within these ancient texts, significantly accelerating locating relevant information. Physical manuscripts are susceptible to deterioration, but digital versions can be stored, backed up, and archived securely for future generations. These can be easily shared online, downloaded, and studied from anywhere worldwide, eliminating the need for physical access to archives or libraries. Researchers can use IndicOCR to conduct comparative studies across different texts or periods.

Digital Inclusion of Local Magazines/Newspapers: IndicOCR can convert printed archives of local magazines and newspapers in different languages into digital formats. It enables the local magazines to be accessible to a broader audience (throughout the country instead of the local region in a state), including individuals with

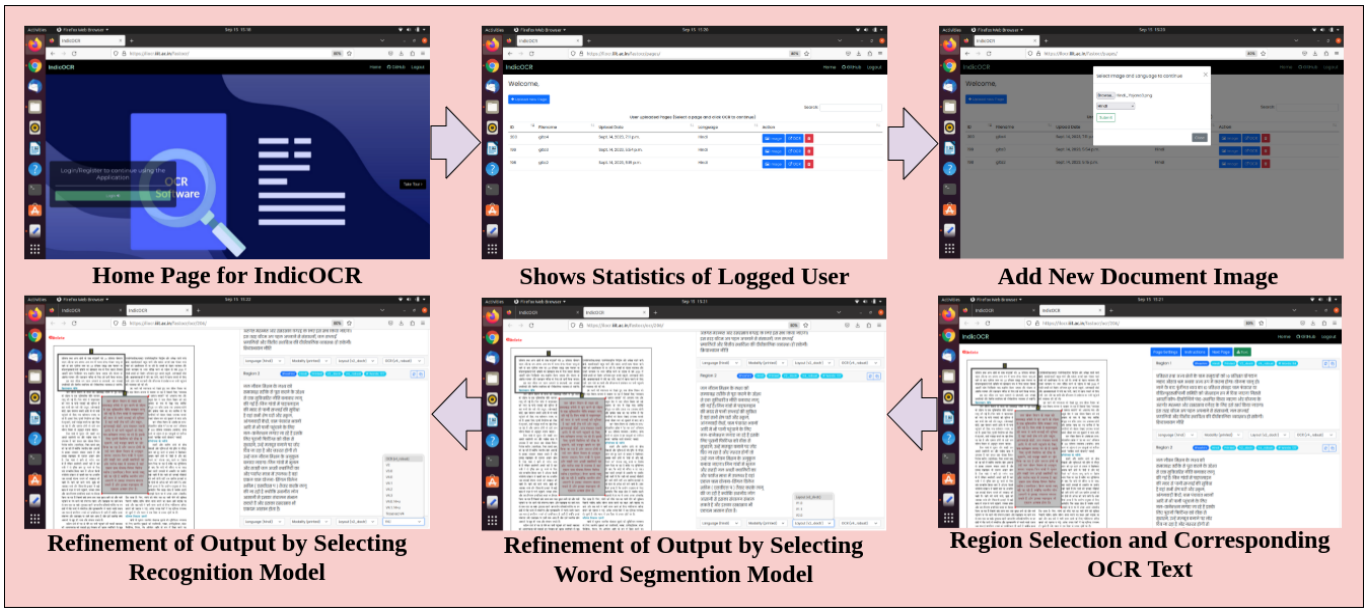

Figure 1: Shows different steps for extracting text from a document image through our IndicOCR pipeline

visual impairments who can use screen readers to access the content. It also cuts down the distribution costs of local magazines. Readers can save time and easily look for specific articles, topics, or keywords of interest as the digitized content is searchable. Local magazines and newspapers often use dialects and regional language variations. IndicOCR helps to preserve these language nuances, contributing to language diversity and heritage preservation.

Digitization of Government Records: Digitizing government records can enhance efficiency, transparency, and accessibility in governmental processes. IndicOCR can convert physical government records into digital formats. It ensures the long-term preservation of essential records and reduces the risk of physical damage or loss. The digitized records become searchable, allowing officials to retrieve specific information from vast archives quickly. It improves decision-making, response times, and data management. Digital records promote transparency and reduce barriers to government employees and the public's easy access to information. It can benefit citizens seeking information on government policies, regulations, or records. Government agencies can use IndicOCR to extract data from documents for analysis, helping identify trends, patterns, and areas that require attention or improvement. It also helps reduce costs, improve disaster recovery, and promote language inclusivity. IndicOCR is a valuable tool for governments looking to streamline operations, improve transparency, and enhance data accessibility.

Education: IndicOCR aids in creating digital educational resources in Indic languages, including textbooks and study materials. It makes accessing, distributing, and updating educational content more manageable for students and teachers. IndicOCR can convert printed materials into accessible formats such as Braille or digital text that works with screen readers while ensuring inclusivity in education for students with visual impairments. The technology is especially relevant during remote learning(such as during a pandemic) by converting printed materials into digital formats for remote distribution.

5 CONCLUSIONS

IndicOCR represents a significant milestone in recognizing India's linguistic diversity. It empowers individuals, organizations, and institutions to work with content in their native languages while preserving linguistic heritage and enabling wider accessibility by providing accurate OCR capabilities for 22 Indian languages.

ACKNOWLEDGMENTS

This work is supported by MeitY, Government of India, through the NLTM-Bhashini project.

REFERENCES

[1] Santhoshini Gongidi and CV Jawahar. 2021. iiit-indic-hw-words: A Dataset for Indic Handwritten Text Recognition. In *International Conference on Document Analysis and Recognition*. Springer, 444–459.
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[3] Mindee. 2021. docTR: Document Text Recognition. https://github.com/mindee/doctr.
[4] Mohamed Ali Souibgui and Yousri Kessentini. 2020. De-gan: A conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 3 (2020), 1180–1191.