

SAMVARDHINI :- 01/06/2025

Volume – 7

Issue – 1

(ISSN ONLINE:- 2583-7176)

<https://samvardhini.in>**Srikanth Venkatesh**

Head, IP Core Services, Tata  
Consultancy Services. Sneha  
Kaushik, IPR Consultant, Tata  
Consultancy Services.

## Exploring new frontiers in AI based OCR techniques for script processing

**Srikanth Venkatesh**

### ABSTRACT

*Optical Character Recognition (OCR) enables digitization of text to a machine-readable text format, wherein the text can be handwritten characters, images or in printed format. With the emergence of Artificial Intelligence (AI)-Machine Learning (ML), OCR has moved from a simple text-recognition tool to a smart text understanding system that can extract, interpret and automate decision making from any type of document. This study explores new frontiers in the OCR technology that has been enabled by the advent of AI/ML by understanding the innovative landscape. The analysis highlights key innovative tools/platforms available across the globe along with details of tools/platforms made in India. Further the patenting trends are examined to gain insights on emerging technologies, the key players enabling the technologies, and filings across the globe.*

### KEYWORDS

*Optical Character Recognition (OCR), Artificial Intelligence (AI)-Machine Learning (ML), AI-based OCR, script digitization*

## **1. Introduction: Technology perspective of AI based OCR**

Converting ancient texts into digital formats makes them more accessible to researchers, scholars and subject matter experts thereby providing avenues for crowdsourcing these historically relevant scriptures. Further, creation of digital archives and online database of manuscripts enhances both accessibility and preservation. Often, degraded or fragmented source documents can also be made amenable for research by utilizing digital technologies. In one such instance, Manuscripts Conservation and Digitization initiative at “The Mythic Society” is worth mentioning. Here the project’s objective was to conserve and digitize palm leaf manuscripts at the Oriental Research Institute (ORI) repository using scientific methods such as AI powered OCR.

AI-powered Optical Character Recognition (OCR) leverages the processing power of artificial intelligence techniques to significantly boost the capabilities, precision and accuracy of conventional OCR systems. AI-powered OCR methods and systems have a proven track record wherein recognition of various languages, understanding document structures, and handling complex layouts can lead to efficient data extraction and document processing capability.

Further the integration of AI with multi-language Natural Learning Process (NLP) has meaningfully expanded the capacity of OCR to accurately extract and interpret text from diverse script as well as language contexts. Deep learning algorithms and Machine learning techniques are deployed to interpret context, understand a document’s structure and then support multi-lingual data processing for text processing environments. Hence NLP based AI models with their inherent capabilities to learn from large datasets of complex scripts and improve accuracy especially with handwritten text, noisy images with interspersed signals and noise, form the foundation elements of next generation agile data processing systems.

Referring to Indic OCR-Language models, there is a lot of research here as the Indic languages are more challenging due to script complexity and historic documentation styles. Kraken OCR and T5 transformers are two most powerful technologies which are being explored for Indic languages especially Sanskrit, wherein these technologies have been used to digitize palm leaf manuscripts, old manuscripts and early printed editions.

Kraken OCR is an OCR system optimized for historical, especially non-Latin script material. The key features of kraken's include a fully trainable layout analysis, reading order and character recognition. The system supports Right-to-Left, BiDi, and Top-to-Bottom. Further the T5 (Text-to-Text Transfer Transformer) is a powerful language model that performs OCR (Optical Character Recognition) post-correction. Instead of being trained

solely for OCR, T5 treats OCR as a text-to-text problem, where the input is a corrupted image and the output is the corrected text. This allows T5 to leverage its strong text processing capabilities for OCR tasks

## **2. Research Work:**

### **2.1 Comprehensive study of existing AI powered OCR Tools**

For digitizing Sanskrit texts using AI based optical recognition systems several specialized platforms and tools have been developed abroad and by Indian institution and researchers. Listed below are curated platforms/tools, wherein section 2.2 covers global tools while the section 2.3 refers to Indian platforms/tools which are made in India.

### **2.2 Existing Global tools in AI powered OCR:**

#### **1. ABBYY FineReader:**

The ABBYY FineReader Engine software development kit utilizes AI OCR to enable software developers to create applications that extract textual information from paper documents, images or displays.

#### **2. Google Cloud Vision API**

The Google Cloud Vision API is configured to identify images from pre-trained models of huge image datasets based on machine learning techniques. The tool further classifies the images into thousands of categories to pick up on objects, places and faces to produces results with a certain confidence value.

#### **3. Amazon Textract:**

Amazon Textract is a machine learning (ML) service to automatically extract text, handwriting, layout elements and data from scanned documents. The tool goes beyond simple optical character recognition (OCR) to identify, understand and extract specific data from documents.

#### **4. Microsoft Azure AI Vision:**

Azure AI Vision is a cloud-based computer vision service that allows developers to analyze images and videos using several advanced algorithms. Azure AI Vision provides various features like image analysis, OCR, object detection and facial recognition, enabling applications to interpret visual content and further extract meaningful insights.

### 5. Tesseract OCR:

**Tesseract** is a popular OCR engine that can operate with various operating systems. The Version 4 includes a Long Short Term memory (LSTM) based OCR engine as well as models for many additional languages and scripts, bringing the total languages supported to 116.

### 6. Nanonets:

Nanonets is an AI-driven solution designed to automate document processing and data extraction workflows. Leveraging advanced concepts of Optical Character Recognition (OCR) along with deep learning models, Nanonets enables organizations to automate document-heavy business processes like accounts payable, order processing and insurance underwriting.

### 7. IronOCR :

IronOCR is a C# and .NET library that specializes in the OCR, barcode and QR code reader technologies. The IronOCR extracts text, barcodes, and QR codes from several image and PDF formats using the latest Tesseract 5 engine.

## **2.3 Existing Indian Tool in AI powered OCR:**

### 1. The Asiatic Society in Kolkata:

- Artificial intelligence (AI) is leveraged to decipher and make extensive collection of ancient manuscripts and historical documents accessible. The documents/manuscripts can be in languages like Sanskrit, Pali and various other Indian scripts.
- This project, part of a broader National Language Learning Initiative, employs OCR technology to analyze script depth and improve transcription accuracy.
- The goal of this project is to make this rich archive available to scholars worldwide, many of whose contents have never been deciphered before.

### 2. IndicOCR: This is a repository containing code for various OCR models tailored for classical Sanskrit document image.

- Includes models such as IndicOCR-V2 CNN RNN for recognizing Sanskrit texts
- Provides tools for computing error rates such as character error rate (CER) and word error rate (WER)

- Supports training and evaluation of custom OCR models.
3. SanskritShala by IIT Kanpur & IIT Kharagpur
    - SanskritShala which is a web-based application allows a user to get real-time analysis for the given input. The tool has a user-friendly interface that supports interactive data annotation features that allow annotators to correct the system predictions when it makes mistakes.
    - The tool facilitates computational linguistic analyses for several tasks including word segmentation, morphological tagging, dependency parsing, and compound type identification
  4. Sanskrit OCR by Auroville:

The Sanskrit OCR online tool developed by Auroville offers a simple browser-based interface for uploading images to obtain OCR results. The tool also provides options to toggle between Google and tesseract OCR engines.

## **2.4 Global Patenting trends for AI based OCR**

With today's evolving technology landscape, patenting trends serve as a crucial parameter of innovation reflecting where the research and development is concentrated on and which technologies are poised to shape up the future.

This section focuses on the patenting trends including key players, technologies at the forefront and geographical shifts in patenting activity which is essential for businesses researchers and policymakers alike.

The patent collection created and analyzed in this section was created in conjunction with Patentscope, USPTO website, InPass, WIPO (World Intellectual Property Organization) database, Lens database as well as licensed/paid databases such as PatSeer and consists of published patent applications and granted patents for the time duration of 2020 to 2025.

### **2.4.1 Key Players:**

Leading the patent filing in AI-based OCR technologies is a mixture of global technology giants and innovation companies across the globe. IBM is a long-standing leader in patents, followed by Google's parent company Alphabet which is especially active with AI based patents. South Korea's Samsung and Japanese Toshiba appear in top ten patent filers for AI led OCR innovations. The Chinese also have a strong presence with Baidu & Ping An Insurance appear in top ten patent filing list.

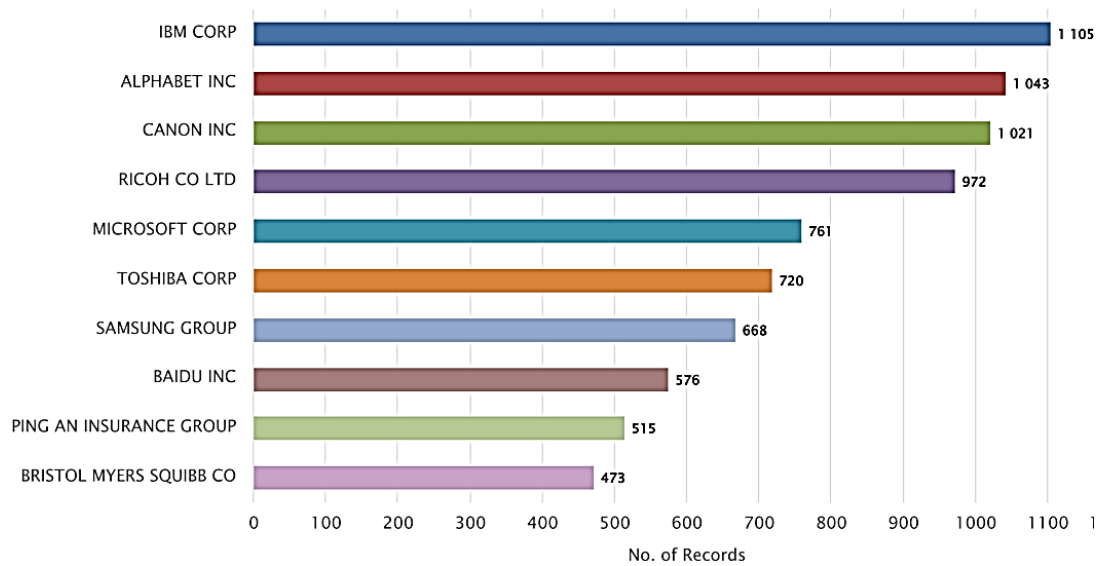


**Exhibit 1: Key players filing maximum patents for AI- based OCR technologies**

Among the Indian innovators, organizations such as Tata Consultancy Services (TCS), Infosys and Wipro are leading IT service companies with significant patent activity in the AI based OCR technology domain.

#### **2.4.2 Key Jurisdictions:**

Owing to flagship programs by the Government of India such as “Viksit Bharat”, “Make in India”, “StartUp India” etc., there has been a surge in inventions & innovations. Hence it can be seen that India is rapidly emerging as a significant player and the applications filed in India showcases the growing indigenous innovation trend. International Filings (WIPO), South Korea, Japan & European Patent Office (EPO) also contribute significantly to the AI based OCR technologies landscape. Rest of the trend is presented below:

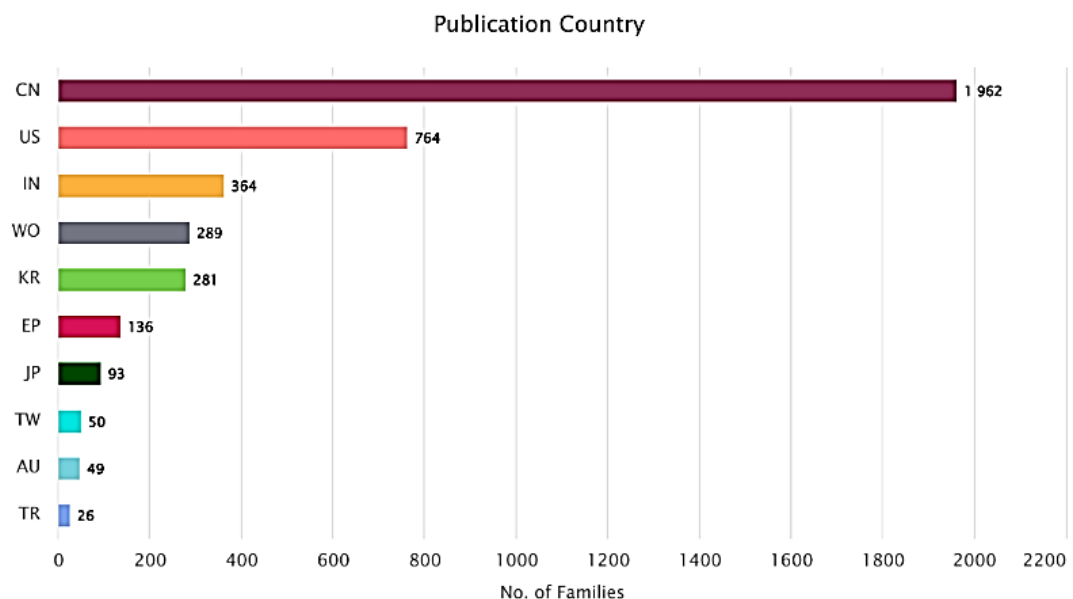


**Exhibit 2: Top filing jurisdictions for AI-based OCR technologies**

### **2.4.3 Filing trends:**

For OCR based technologies - The past five years have seen a consistent upward trajectory in global patent filing. This surge may have happened due to the rapid increase in the usage of artificial intelligence and natural learning process-based technologies.

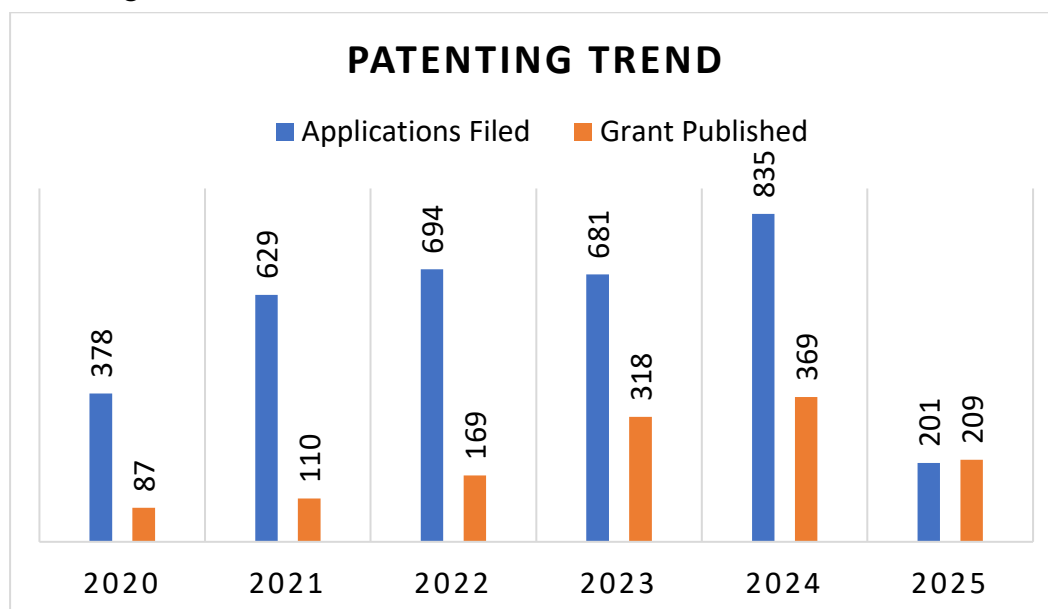


Exhibit 3: Patent filing trends for AI-based OCR technologies

### **2.4.4 Technology specific Patenting trends:**

The patent landscape has seen a rapid exponential growth with the emergence of Artificial Intelligence (AI) along with Machine learning (ML) and Natural Language Processing (NLP), wherein the existing trends focus on AI/ML algorithms, ML models, training test data for the ML models and electronic data with focus on document image. There is also growing interest in user interface



Exhibit 4: Technology specific filing trends for AI-based OCR technologies

#### 2.4.5 Key Patents filed for AI-based OCR technologies:

This section lists out a few interesting patents filed across the globe for AI-based OCR technologies.

| Patent No and Title                                                                                                        | Assignee            | Brief Description of the patent                                                                                                                                                                                                                                                                                                                                                                                                        |
|----------------------------------------------------------------------------------------------------------------------------|---------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <a href="#">US2023094787A1: Utilizing machine-learning based object detection to improve optical character recognition</a> | ADOBE INC (US)      | The present invention relates to systems, methods for accurately enhancing optical character recognition with a machine learning approach. The invention determines words from reverse text, vertical text, and a typically sized text. An optical character recognition model is utilized to determine words from the modified versions of the reverse text glyphs, the vertical text glyphs, and/or the typically sized text glyphs. |
| <a href="#">KR102561878B1 : AI OCR ai blue OCR reading system and method based on machine learning</a>                     | BELLTECHSOFT CO LTD | Key inventive steps of this patent includes the simultaneous processing of multiple-page documents, the application of a learning model derived from machine learning to analyze and categorize text areas, and the integration of a character correction unit that refines recognized characters                                                                                                                                      |

|                                                                                                                          |                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|--------------------------------------------------------------------------------------------------------------------------|---------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <a href="#"><u>KR102496620B1: AI OCR AI-based search function and OCR electronic research note management system</u></a> | KMIEG INC (KR)            | Key inventive steps include a user input unit for entering various types of research data, a supplementary information generating unit that predicts and acquires relevant data, and a research note information processing unit that classifies and stores this information. The system also features a platform server that connects with external agency servers to provide additional data, enhancing the research note's context.                      |
| <a href="#"><u>US2022180113A1: Method and system for document classification and text information extraction</u></a>     | Tata Consultancy Services | This patent by TCS is a multistage image pre-processing approach for efficient, effective, and accurate background noise removal from the classified document. The techniques used includes unsupervised clustering, identification, segmentation, masking, contour approximation, selective subtraction, and dynamic thresholding.                                                                                                                         |
| <a href="#"><u>CN113168527A: Systems and methods for extracting information from a physical document</u></a>             | Google Inc                | The method of this Google patent extracts a set of candidate labels by performing a search for the candidate labels. The method disclosed includes determining a canonical label for each annotated value based on at least in part on the associated candidate labels, and mapping at least one annotated value to an action that is presented to the user, which is based on at least in part on the canonical label associated with the annotated value. |

### **3. Conclusion:**

Based on the study/survey it can be concluded that the integration of AI and NLP have significantly advanced the traditional optical character recognition by transforming it from simple character recognition tool into an intelligent text analyzing system. The usage of these models is not just limited to modern printed texts/documents but have also been widely used for analysis of ancient scripts. Further ahead, OCR is bound to be more autonomous, dynamic and adaptive in many more languages and will also be programmed to handle ancient, and regional scripts for not just recognizing text but also

to extract meaningful structure and relationship in the context, thus playing a pivotal role in cultural preservation, academic research and digital accessibility across languages and disciplines.

#### 4. Bibliography:

1. Tools used for patent related information:

Patseer - <https://patseer.com/> - Date: 04/06/2025, Time: 11:00 AM

ESpaceNet - <https://worldwide.espacenet.com> – Date:02/06/2025, Time: 12:00 PM

PatentScope WIPO - <https://patentscope.wipo.int/search/en/search.jsf> Date: 02/06/2025, Time: 1:50 PM

INPass - <https://iprsearch.ipindia.gov.in/PublicSearch/> -Date:02/06/2025, Time: 3:00 PM

USPTO Search - <https://www.uspto.gov/patents/search> – Date:02/06/2025, Time: 4:00 PM

Lens - <https://www.lens.org/> - Date: 02/06/2025, Time: 9:00 PM

2. <https://mythicsociety.org/> - Date: 27/05/2025, Time: 11:25AM

3. <https://sanskritshala.github.io/> Date: 25/05/2025, Time: 10:25AM

4. <https://sri.auroville.org/> Date: 26/05/2025, Time: 12:25PM.

5. <https://timesofindia.indiatimes.com/city/kolkata/asiatic-society-launches-ai-initiative-to-decipher-ancient-manuscripts-in-archives/articleshow/120171417.cms>  
Date: 28/05/2025, Time: 2:25PM.

6. <https://indic-ocr.github.io> Date: 28/05/2025, Time: 10:45PM.

7. <https://www.curvestone.io/blog-post/comparison-of-optical-character-recognition-ocr-services-from-microsoft-azure-aws-google-cloud> Date: 28/05/2025, Time: 3:20 PM.

8. <https://github.com/mittagessen/kraken> Date: 02/06/2025, Time: 11:50 AM

9. <https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/overview-ocr>  
Date: 30/05/2025, Time: 4:50 PM