OCR Services Evaluation Report

Patrik Baldon
patrik.baldon@studenti.unipd.it

Felice Francario
felice.francario@studenti.unipd.it

Abstract

This project explores the implementation and evaluation of various Optical Character Recognition (OCR) technologies, including Tesseract OCR, OCR.Space, Azure OCR, Google Cloud OCR, and AWS Textract. The objective is to analyze their performance in extracting textual data from images, focusing on accuracy, processing speed, and cost-effectiveness. Each service is examined based on its processing type, supported formats, and pricing structure. The analysis aims to provide insights into the strengths and weaknesses of each solution, guiding users in selecting the most appropriate OCR tool for their specific needs. Special emphasis is placed on accuracy comparisons using data from a variety of sources. The findings contribute to a better understanding of the state of the art OCR services capabilities in various application scenarios, such as document digitization, text recognition in natural scene images, automated data entry from handwritten documents, and multilingual recognition.

1. Introduction

In recent years, Optical Character Recognition (OCR) technology has seen significant advancements, enabling the automated extraction of textual information from a variety of image-based sources. OCR systems are widely used across multiple domains, including document digitization, automated data entry, and accessibility enhancement for visually impaired individuals. The increasing reliance on OCR solutions necessitates a thorough evaluation of their performance across different scenarios to ensure optimal selection for various applications.

This project aims to evaluate and compare five leading OCR systems: Azure OCR, Tesseract OCR, AWS Textract, OCR.Space, and Google Cloud Vision. Each of these solutions leverages different underlying technologies, including deep learning-based approaches such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks, as well as heuristic-based methods for text recognition. The primary objective is to assess their effectiveness in terms of accuracy and processing efficiency, scalability, and cost-effectiveness when applied to various types of textual data, such as printed documents, handwritten text, and complex scene images.

The evaluation process involves rigorous benchmarking using diverse datasets that encompass multiple languages, font styles, and document formats. Key performance indicators (KPIs) such as character recognition accuracy, word accuracy, processing speed, robustness under various scenarios will be analyzed to provide a comprehensive understanding of each system's strengths and limitations.

Through this comparative analysis, we aim to identify the most reliable and cost-efficient OCR solution that aligns with specific application requirements, such as real-time processing, cloud deployment, or on-premise implementations.

2. Related Work

The evaluation of Optical Character Recognition (OCR) systems has been an active area of research, with numerous studies comparing the performance of various OCR services using different datasets and evaluation metrics. Two notable studies that provide a comprehensive analysis of OCR solutions are discussed below.

The study conducted by Curvestone [10] compares OCR services provided by Microsoft Azure, AWS Textract, and Google Cloud Vision. Their evaluation focuses on metrics such as Intersection over Union (IoU) for word bounding boxes and the correctness of detected words across different document types. The findings indicate that cloud-based services generally provide higher accuracy in structured documents, while local solutions may struggle with layout complexity.

Similarly, the OCR benchmarking report by AI Multiple [4] analyzes the accuracy levels of major OCR solutions, including Tesseract, Google Cloud Vision API, and AWS Textract, across various document types. The benchmark results reveal that while printed text can achieve over 95% accuracy across all solutions, handwritten text and complex layouts present significant challenges. This study recommends Tesseract for cost-effective solutions when dealing with printed documents and cloud-based services for more

complex requirements.

Despite the valuable insights provided by these studies, our approach introduces several key differences that enhance the comprehensiveness of OCR evaluation.

Our evaluation employs a more comprehensive set of performance metrics, including the Levenshtein distance, Word Accuracy, and Character Error Rate (CER). These metrics provide a granular assessment of OCR performance by quantifying the degree of text recognition errors at both the character and word levels, ensuring a more detailed and objective comparison. In contrast, previous studies have primarily focused on general accuracy without considering the specific error patterns and their implications for downstream applications. Our evaluation also considers various possible ocr scenarios, including 6 different language scripts, which will be expanded on in the next section.

By adopting a holistic evaluation framework and a diverse dataset of real-world documents, our project aims to provide deeper insights into the practical applicability of OCR systems in various scenarios.

3. Datasets

The datasets were filtered to image dimensions of less than 1MB due to operational constraints of the free subscription plan of OCR.Space and to ensure a fair comparison between the services. The following are the datasets that were used in evaluating the OCR services:

Focused Scene Text 2013-2015

- **Source and size of datasets:** The dataset consists of 233 images collected from various scene environments to benchmark scene text recognition algorithms[6]. For our evaluation the first 150 samples, satisfying the 1 MB maximum size threshold, were considered.
- **Characteristics:** The images contain English text captured in diverse lighting and perspective conditions, simulating real-world challenges for OCR systems.

SROIE 2019

- **Source and size of datasets:** The SROIE (Scanned Receipts OCR and Information Extraction) 2019 dataset[8] includes over 1000 scanned receipts, specifically designed for structured data extraction and OCR evaluation. For our evaluation the first 150 samples, satisfying the 1 MB maximum size threshold, were considered.
- **Characteristics:** This dataset primarily consists of financial documents containing English text with a variety of font styles, alignments, and noise levels, which present challenges for OCR systems.

Figure 1. Example of receipt document images in the SROIE dataset

IAM Handwriting Database

- **Source and size of datasets:** The IAM Handwriting Database consists of approximately 1500 scanned handwritten pages, containing more than 100,000 words written in English, collected from diverse sources[5]. For our evaluation 150 samples, satisfying the 1 MB maximum size threshold, were considered.
- **Characteristics:** The dataset features a variety of handwriting styles, ink types, and document layouts, making it a valuable resource for training and evaluating handwriting recognition models.

ICDAR 2019 Robust Reading Challenge on Multilingual Scene Text Detection and Recognition

- **Source and size of datasets:** This dataset consists of thousands of images collected from real-world scenes containing text in multiple languages, including English, Chinese, Japanese, Korean, Hindi and Arabic[7]. For our evaluation the first 50 sample images of each language, satisfying the 1 MB maximum size threshold, were considered.
- **Characteristics:** The dataset provides a challenging environment for OCR systems due to its multilingual content, diverse fonts, varying lighting conditions, and occlusions commonly found in natural scenes.

4. Main caracteristichs of each OCR service

Processing Type

OCR systems can be classified based on their deployment model. **Tesseract OCR** is a local solution that processes documents on the user's machine, offering enhanced privacy and control over data. In contrast, cloud-based solutions such as **OCR.Space, Azure OCR, Google Cloud OCR, and AWS Textract** process documents on remote servers, providing scalability, accessibility, and higher computational power, but may raise concerns regarding data privacy and latency.

Costs and subscription plans

Cost is a significant factor when choosing an OCR solution. **Tesseract OCR** is open-source and free to use, making it a cost-effective option for local processing. **OCR.Space** offers a free subscription plan with limits of 2500 requests/month and a file size limit of 1 Mb. No credit card information is required to sign up for the free tier. To exapand on the size and requests limits, it also offers various subscription models starting from $30 per month[9].**Azure OCR** offers a free subscription plan(with some functional limitations) of 12 months consisting of 5000 free transactions a month after which it charges from $0.4 to $1.00 per 1,000 requests based on the volume of requests. It also offers the option of 200$ in credit to be used in 30 days[2]. **Google Cloud OCR** offers 1000 requests a month for free and then costs$1.50 per 1,000 requests, although also providing $300 worth of credits for new users in the first 3 months[3].**AWS Textract** offers a 3 month free tier plan for new users with a limit of 1000 pages/month after which the price becomes $1.50 per 1,000 pages for standard operations[1].

T

4.1. Properties and Limitations

Despite their advanced capabilities, the evaluated OCR solutions exhibit various limitations that can impact their effectiveness in different use cases.

Tesseract OCR is a powerful open-source tool; however, it requires extensive preprocessing to achieve optimal accuracy. Without appropriate image enhancement, such as noise reduction and contrast adjustment, its performance deteriorates significantly. Additionally, it struggles with cursive handwriting and non-standard fonts, making it less suitable for processing complex or handwritten documents. This OCR can't handle multilingual text, restricting its application only to the english language. Another major drawback is its lack of cloud integration, limiting scalability and accessibility for large-scale or distributed applications.

OCR.Space, being a cloud-based service, offers ease of access but comes with significant constraints. The free tier imposes strict file size limitations of up to 1 Mb per document, and up to 5 Mb with the cheapest monthly subscription plan of 30$. It also has a limit of 3 pages per pdf, making it unsuitable for processing large documents. Moreover, its processing speed is inconsistent and largely dependent on server load, which can introduce delays in high-demand scenarios and affect the efficiency of time-sensitive applications.

Azure Computer Vision OCR provides robust text recognition capabilities and has an extremely economically convenient free tier which unfortunately comes with several limitations: processing only the first 2 pages of pdf documents and restricting the file size of images to 4 Mb. With the paid version, the limits increase to 2000 pages and 500 Mb respectively. The requirement of an active internet connection for cloud processing poses challenges for offline applications, making it unsuitable for environments with limited or unreliable connectivity.

Google Cloud Vision OCR excels in text recognition but has noticeable limitations when handling complex table layouts, often resulting in inaccurate text extraction from structured documents. It has a limit of quota limit 1800 requests per minute, but can be modified by submitting a special request to Google. Furthermore, it has a relatively low maximum image file size limit of 20MB regardless of tier, which can limit very high resolution images but is more than plenty for any normal image.

AWS Textract is well-suited for structured document analysis; however, it is among the most expensive solutions, making it a less viable option for budget-conscious users. The processing speed for large document batches tends to be slower compared to other cloud-based services, potentially affecting workflow efficiency. Handling datasets in the AWS cloud environment is also less straightforward to use and more time consuming for beginners compared to the other services. For synchronous operations, JPEG, PNG, PDF, and TIFF files have a limit of 10 MB in memory, while PDF and TIFF files have a limit of 1 page. For asynchronous operations, JPEG and PNG files still have a limit of 10 MB in memory while PDF and TIFF files have a limit of 500 MB in memory. This significantly limits AWS Textract for High-resolution images or synchronous operations dealing with multi page documents. This OCR also can't handle many languages, restricting its application only to variants of latin characters.

These limitations highlight the importance of selecting the appropriate OCR solution based on specific project requirements, taking into account factors such as document size, complexity, processing speed, cost, and security considerations.

5. Evaluation Method

The evaluation process began with selecting the best and most diverse available ground-truth datasets to ensure a reliable comparison of OCR services. Since the format of these datasets varied, customized preprocessing code was developed for each dataset to standardize the ground-truth data and facilitate implementation. Once the datasets were prepared, each OCR service processed a selection of files, extracting text from them. The extracted text was then evaluated using multiple metrics to measure the accuracy and effectiveness of each model. Additionally, to assess the potential for post-processing improvements, the same metrics were applied to model outputs refined by a large language model, Gemini 1.5 Flash. This step provided insights into which OCR models retained the most information and demonstrated the highest potential for enhancement. To determine the best-performing OCR service for each dataset type, a scoring system was established. Weights were assigned to different evaluation metrics based on their relative importance, allowing for a prioritized assessment of OCR models. The following section gives a detailed look at all the metrics used for the evaluation of the models.

5.1. Evaluation Metrics

Levenshtein Accuracy

The Levenshtein accuracy measures the similarity between the extracted text and the ground truth text by computing the number of single-character edits (insertions, deletions, substitutions). The formula used is the following:

$$\left(1 - \frac{\text{Levenshtein Distance}}{\max(\text{len}(\text{Extracted Text}), \text{len}(\text{Ground Truth}))}\right) \times 100$$

The Levenshtein distance is a string metric used to measure the difference between two sequences of characters. It is defined as the minimum number of single-character edits required to transform one string into another.

Word Accuracy

This metric evaluates the proportion of correctly recognized words by comparing the OCR-extracted words with the ground truth:

$$\left(\frac{\text{Number of Matching Words}}{\text{Total Words in Ground Truth}}\right) \times 100$$

Character Error Rate (CER)

CER quantifies the percentage of incorrect characters in the OCR output:

$$\left(\frac{\text{Levenshtein Disance}}{\text{Total Characters in Ground Truth}}\right) \times 100$$

Score

The overall performance is quantified using a combined score, which considers the Levenshtein accuracy, word accuracy, and character error rate (CER). The score is computed as a weighted sum of these metrics, where the weights are determined by the parameters α, β, and γ. The formula for calculating the score is as follows:

$$\text{Score} = \alpha \times \text{Levenshtein Acc.} + \beta \times \text{Word Acc.} - \gamma \times \text{CER}$$

where: α controls the contribution of Levenshtein Accuracy, β controls the contribution of Word Accuracy, and γ controls the penalty applied to the Character Error Rate.
This combined score offers a balanced measure that accounts for both accuracy and error, with the ability to fine-tune the influence of each metric based on specific application requirements.
The optimal values for these parameters, $\alpha = 1.3$, $\beta = 0.6$, and $\gamma = 1.5$, were chosen based on empirical evaluation of the metrics, giving more priority to the complete detection of the text present in the images while also considering the accuracy of the detected words.

6. Results

The results of our experiments indicate that cloud-based OCR services outperform local-based ones in terms of processing speed. Among all tested services, Google Cloud Vision OCR was the fastest, with an average processing time of 0.60 seconds per image. Other cloud-based services had processing times ranging from 1.5 to 2 seconds, while the slowest was the open-source Tesseract, which averaged 2.60 seconds when run on Google Colab. However, it is important to note that OCR.space exhibited instability when processing large batches of files, frequently encountering unpredictable server timeouts, which could impact reliability in high-volume applications.

The following sections provide a detailed breakdown of OCR performance across different evaluation scenarios.

6.1. Natural scene Images

Using the Focused Scene Text 2013-2015 Dataset, Azure achieved the best overall performance, followed closely by OCR.space, as shown in Table Table 1. As expected, Tesseract struggled significantly, failing to accurately extract text from natural environments. Applying LLM-based output refinement (Gemini 1.5 Flash) did not improve results and, in fact, slightly decreased overall accuracy. This is likely

because natural scene images typically contain a small set of highly specific words (e.g., store names, billboards, and signage).

For natural scene images Azure ends up being the best service by all metrics, with OCR.Space being a close second. Google provides the best word accuracy, indicating that its the least likely to provide misspelled words but the overall score indicates that,in comparison, it has a harder time with identifying all the text present in an image.

API	Levenshtein Accuracy (%)	Word Accuracy (%)	Character Error Rate (%)	Average Score
Azure	**84.84**	**83.33**	**18.26**	**132.89**
OCR.space	84.15	79.85	19.60	127.90
AWS	76.60	78.45	31.14	99.95
Google	79.40	86.78	41.30	93.35
Tesseract	15.69	14.59	88.72	-103.92

Table 1. Performance metrics of various OCR APIs on a scene image dataset.

6.2. Handwriting Images

Using the IAM Handwriting Dataset, Tesseract again is clearly the worst service. Indicating its incapability in handling handwritten text. Just like the previous case, as can be seen in table Table 2, Azure OCR is clearly the best according to all the evaluation metrics and the other services follow it showing a very close behavior. AWS Textract is a close second, beating Google in every metric except word accuracy. The third overall best model is Google's cloud vision OCR, again showing impressive word accuracy. However, it is interesting to note that in this case, refining the outputs through an LLM provides a significant impact on the OCR.space model, increasing its overall performance score to 107.8 ahead of Google's regular score. This indicates the potential in improving ocr.space's model through LLM integration, limiting its frequent misspellings.

API	Levenshtein Accuracy (%)	Word Accuracy (%)	Character Error Rate (%)	Average Score
Azure	**82.03**	**78.03**	**19.38**	**124.37**
AWS	79.97	72.01	21.78	114.49
Google	76.61	73.92	24.82	106.71
OCR.space	78.25	63.98	22.83	105.86
Tesseract	33.22	13.10	71.33	-55.95

Table 2. Performance metrics of various OCR APIs on handwritten text.

6.3. Documents

For Document text extraction, AWS Textract proved to be the best service by all metrics with Azure OCR being a close second. It is important to note that all the models had a relatively low word accuracy on this dataset possibly due to the misidentification on what constitutes a word in raw unpreprocessed receipts with an abundance of numbers, as can be seen in fig Figure 1. Since all the models face the same playing conditions, this still provides an excellent way to compare the different models in their handling of raw unpreprocessed financial documents. While AWS Textract and Azure OCR, proved to be head and shoulders above the rest, this was the first case in which the open source Tesseract model provided a decent performance even slightly surpassing OCR.space in terms of Levenshtein accuracy. For this case, LLM output refinement had no significant impact on any of the models performances.

API	Levenshtein Accuracy (%)	Word Accuracy (%)	Character Error Rate (%)	Average Score
AWS	**81.98**	**44.23**	**39.66**	**73.62**
Azure	80.21	43.13	42.13	66.96
Google	70.70	43.82	51.25	41.33
OCR.space	65.74	40.25	54.83	27.37
Tesseract	66.34	30.84	55.28	21.88

Table 3. Performance metrics of various OCR APIs on documents

6.4. Multilingual scene Images

To assess multilingual capabilities, each OCR service was tested on five diverse language scripts: Arabic, Chinese, Korean, Japanese, and Hindi. AWS Textract and Tesseract were excluded from this evaluation due to their inability to process these languages effectively. Only three OCR services—Azure, Google Cloud Vision, and OCR.space—were able to process the dataset, as the remaining services struggled with the complexity of these scripts. Word accuracy was used as the sole evaluation metric since other metrics proved inadequate due to variations in reading direction (e.g., right-to-left or top-to-bottom text orientation).Additionally, the ground truth documents did not have text in a sequential order, making word accuracy the most reliable evaluation method. Among the evaluated OCR services, in 4 of the 5 languages, **Azure achieved the highest overall accuracy**, followed by Google Cloud Vision, while OCR.space consistently performed the worst. OCR.Space particularly struggled with Japanese and Chinese text. However, for Hindi, Google Cloud Vision outperformed all other services, achieving the highest word accuracy, while OCR.space was excluded due to its inability to process Hindi text. The detailed results for each language are presented in the tables below.

API	Word Accuracy(%)
Google	**77,34**
Azure	69,01

Table 4. Evaluation on Hindi images

API	Word Accuracy(%)
Azure	**76,80**
Google	71,66
OCR.space	62,00

Table 5. Evaluation on Arabic images

API	Word Accuracy(%)
Azure	**69,07**
Google	63,84
OCR.space	45,34

Table 6. Evaluation on Chinese Images

API	Word Accuracy(%)
Azure	**65,82**
Google	64,47
OCR.space	60,13

Table 7. Evaluation on Korean images

API	Word Accuracy(%)
Azure	**67,00**
Google	66,50
OCR.space	30,99

Table 8. Evaluation on Japanese images

7. Conclusion

Based on these results, we conclude that the best OCR service depends on the specific task. In most cases, **Azure** offers the highest extraction quality and the most versatile performance across different types of text, making it the **best overall choice**.It excels in structured documents, natural scene text, and handwriting recognition, while also providing the best multilingual support, outperforming all other services in language diversity.

Google Cloud Vision OCR follows closely behind Azure in terms of language flexibility and general text recognition, offering a transparent and flexible pricing model. However, when it comes to structured document extraction, **AWS Textract** slightly outperforms Azure, thanks to its specialized tools for handling complex layouts, tables, and forms.

That said, AWS Textract has a major drawback: it cannot handle languages beyond Latin scripts. This severely limits its usability for multilingual OCR tasks. Additionally, while it provides enterprise-level security and seamless integration with other AWS services, it requires files to be uploaded to an Amazon S3 bucket—a cloud storage service. This makes it less beginner-friendly, more time-consuming for large-scale processing, and requires more complex code compared to other services. While beneficial for companies needing secure, scalable workflows, it may not be ideal for users prioritizing speed, multilingual capabilities, and ease of use.

Tesseract also suffers from the same language limitation as AWS, as it struggles with non-Latin scripts and performs poorly on handwriting and natural scene text. It remains viable for structured document recognition, but requires extensive image preprocessing to achieve reliable results.

OCR Space is a cost-effective alternative, particularly for natural scene text and handwriting recognition. However, its 1MB per image limit on the free plan and lower accuracy for structured documents may limit its usability. Interestingly, its accuracy improves when paired with an LLM (Gemini 1.5 Flash), making it a budget-friendly AI-enhanced option.

Future research could explore the capabilities of AI integration in OCR models and determine whether services like OCR Space can improve their accuracy and narrow the gap with more sophisticated models.

References

[1] AWS. Amazon textract pricing. https://aws.amazon.com/it/textract/pricing/, note = Accessed: 2025-01-24.

[2] Azure. Azure ai vision pricing. https://azure.microsoft.com/en-us/pricing/details/cognitive-services/computer-vision/#pricing. Accessed: 2025-01-24.

[3] Google Cloud. Cloud vision prices. https://cloud.google.com/vision/pricing?hl=it#cloud-vision-pricing, note = Accessed: 2025-01-24.

[4] Cem Dilmegani. Ocr benchmarking: Text extraction / capture accuracy ['25], 2025.

[5] fki. Iam handwriting database. https://fki.tic.heia-fr.ch/databases/iam-handwriting-database.

[6] ICDAR. Focused scene text. https://rrc.cvc.uab.es/?ch=2.

[7] ICDAR. Icdar 2019 robust reading challenge on multilingual scene text detection and recognition. https://rrc.cvc.uab.es/?ch=15&com=introduction.

[8] ICDAR. Icdar 2019 robust reading challenge on scanned receipts ocr and information extraction. https://rrc.cvc.uab.es/?ch=13.

[9] OCRSpace. Ocr api prices. https://ocr.space/ocrapi, note = Accessed: 2025-01-24.

[10] Curved Stone. Comparison of optical character recognition (ocr) services from microsoft azure, aws, google cloud. 2022.