

# OCR SERVICES EVALUATION

VISION AND COGNITIVE SYSTEMS  
PROJECT





# INTRODUCTION TO OCR TECHNOLOGIES

## UNDERSTANDING OPTICAL CHARACTER RECOGNITION(OCR):

- **Definition and Purpose:** OCR enables the automated extraction of textual information from images and documents, widely used in digitization, accessibility, and automated data entry.
- **Objectives:** This study evaluates five OCR services-Azure OCR, Tesseract OCR, AWS Textract, OCR.Space and Google Cloud OCR- on multiple ocr scenarios including scene , document, handwritten and multilingual texts.



# OCR (OPTICAL CHARACTER RECOGNITION)

## HOW WORK?

- **Acquisition:** Scanning/photos.
- **Pre-processing:** Filtering, deskewing.
- **Segmentation:** Lines, words, characters.
- **Recognition:** Pattern matching, neural networks.
- **Post-processing:** Dictionary/context checks.
- **Output:** Searchable/Editable formats.



# **OVERVIEW OF EVALUATED OCR SERVICES**





# AZURE OCR

- Cloud-based service, robust text recognition capabilities
- extremely economically convenient free tier 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.
- Can handle many languages



# AWS TEXTTRACT

- Cloud based service, well-suited for structured document analysis
- Among the most expensive solutions, making it a less viable option for budget-conscious users.
- 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.
- It can't handle many languages, restricting its application only to variants of latin characters.



# GOOGLE CLOUD VISION

- Cloud based service
- 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
- Can handle many languages



# TESSERACT

- powerful open-source tool
- requires extensive preprocessing to achieve optimal accuracy.
- struggles with cursive handwriting and non-standard fonts, making it less suitable for processing complex or handwritten documents.
- can't handle multilingual text, restricting its application only to the english language.
- lack of cloud integration, limiting scalability and accessibility for large-scale or distributed applications.





# OCR SPACE

- Free online OCR service, no credit card registration required
- 1 Mb file limit and 3 page PDF limit
- 2500 requests/month
- Supports multiple languages
- Limits are upgradable to 5 Mb and 300,000 requests/month with a PRO subscription



# COST CONSIDERATIONS

- **Tesseract ocr(Free):** Open-source and completely free, making it the most cost-effective but limited in performance.
- **OCR.Space:** Free tier with 1 Mb file limit and 2500 requests/month; paid plans start at \$30/month
- **Azure OCR:** 12-month free tier with 5000 free transactions per month; paid plans range from \$0.4 to \$1 per 1000 requests.
- **Google Cloud Vision OCR:** 1000 free requests per month, then \$1.50 per 1000 requests, with a \$300 free credit for new users.
- **AWS Textract:** Three-month free tier(1000 pages/month), then \$1.50 per 1000 pages; highest cost among the services.



# DATASETS





# Datasets used in OCR Evaluation

- **Focused scene Text 2013-2015:**
  - 150 images collected from various scene environments to benchmark scene text recognition.
  - English text captured in diverse lighting and perspective conditions, simulating real-world challenges for OCR systems.
- **SROIE 2019 (Scanned Receipts OCR and Information Extraction):**
  - scanned receipts, specifically designed for structured data extraction and OCR evaluation. financial documents
  - English text with a variety of font styles, alignments, and noise levels, which present challenges for OCR systems
- **IAM Handwriting Database:**
  - scanned handwritten pages, containing more than 100,000 words written in English, collected from diverse sources
  - variety of handwriting styles, ink types, and document layouts
- **ICDAR 2019 Multilingual Dataset:**
  - real-world scenes containing text in multiple languages, including English, Chinese, Japanese, Korean, Hindi and Arabic



# EVALUATION





# EVALUATION METHODOLOGY

- Processing types- speed
- Accuracy Metrics
- LLM Post-Processing



# Evaluation Metrics

## 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).

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



# Evaluation Metrics

## WORD ACCURACY:

Evaluates the proportion of correctly recognized words.

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





# Evaluation Metrics

CHARACTER ERROR RATE:

Quantifies the percentage of incorrect characters.

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



# Evaluation metrics

SCORE:

Different weights to give greater importance to the levenshtein distance.

Word accuracy with inefficient ground truth.



# LLM Post-Processing

- GEMINI 1.5 FLASH LLM Model was used on outputs to potentially verify the impact of AI integration with the various models
- Corrects spelling mistakes
- Provides an idea of which services have the most potential to be customly upgradeable and the overall information outputted by a model



# Prompt Template

```
prompt_template='''You are an advanced language model specialized in text correction. You will receive text extracted from images using OCR (Optical Character Recognition) models. Yo

### Instructions:
1. Correct spelling errors: Replace any misspelled words with the correct ones.
2. Maintain context: Ensure that the corrected text aligns with the overall meaning and structure of the original input.
3. Handle OCR-specific errors:
  - Fix common OCR mistakes such as incorrect substitutions of similar-looking characters (e.g., "rn" misread as "m").
  - Handle mixed-case errors, such as "tHiS is" to "This is."
4. Do not alter proper nouns, numbers, or special characters unless they are obviously incorrect.
5. Do not add or remove punctuations

Output the corrected text clearly and concisely.

### Example Inputs and Outputs:

**Input:**
This 1s an exarnple of OCR t3xt.

**Output:**
This is an example of OCR text.

**Input:**
Th@ rn0del will c0rrect err0rs.

**Output:**
The model will correct errors.

Focus on accuracy and consistency. Your goal is to produce clean and coherent text that closely resembles the intended content. Only output the corrected text and noting else'''
```

# RESULTS



# PROCESSING SPEED

- **Google Cloud Vision** on average is by far the fastest service
- **Tesseract** is the slowest
- **OCR.Space** performed well, but was the only service with server instability when processing batches of images

API	Google	Azure	OCR.Space	AWS	Tesseract
Average speed (s)	<b>0.60</b>	1.69	2.07	2.15	2.60

# SCENE TEXT RECOGNITION





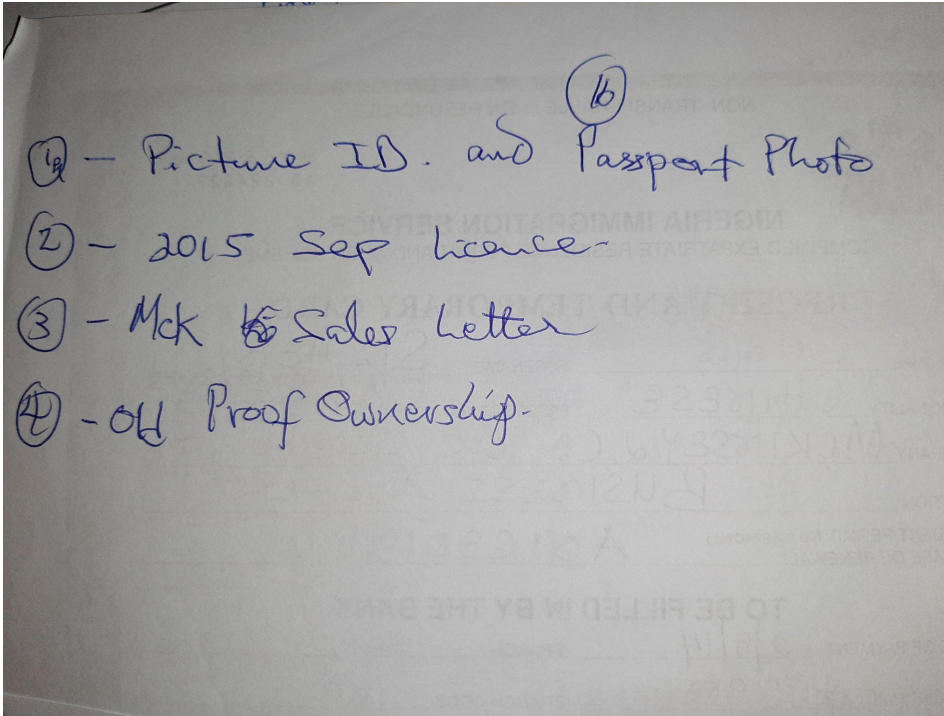
# SCENE TEXT RECOGNITION RESULTS

- **BEST OVERALL PERFORMER: AZURE OCR**
- **Google cloud vision** excelled in word accuracy but likely struggled with identifying all words present in complex scenes, or added unnecessary text
- **Tesseract** failed to accurately extract scene text, performing the worst among all services.

API Name	Levenshtein Accuracy(%)	Word Accuracy(%)	Character Error Rate(%)	Average Score
<b>Azure OCR</b>	<b>84.84</b>	83.33	<b>18.26</b>	<b>132.8941</b>
OCR.space	84.15	79.85	19.59	127.9055
AWS Textract	76.60	78.45	31.14	99.95072
Google	79.40	<b>86.78</b>	41.30	93.34682
Tesseract	15.69	14.58	88.72	-103.92



# HANDWRITING OCR

- 
- A handwritten list of four items on a piece of paper. The paper has faint, mirrored text from the reverse side, including "NIGERIA IMMIGRATION SERVICE", "EXPATRIATE RESIDENCE PERMIT", and "TO BE FILLED IN BY THE BANK". The list is written in blue ink and numbered 1 through 4. A circled 'b' is written above item 1.
- ① - Picture ID. and <sup>(b)</sup> Passport Photo
  - ② - 2015 Sep licence
  - ③ - Mck ~~to~~ Sales Letter
  - ④ - OH Proof Ownership.



# HANDWRITING OCR RESULTS

- **BEST PERFORMER: AZURE OCR** achieved the highest accuracy across different handwriting styles and ink types.
- Tesseract struggled significantly, making it unsuitable for handwritten documents.

API	Levenshtein Accuracy (%)	Word Accuracy(%)	Character Error Rate (%)	AVG Score	AVG score with LLM
<b>Azure OCR</b>	<b>82.03</b>	<b>78.03</b>	<b>19.39</b>	<b>124.3798</b>	<b>114.5709</b>
AWS Textract	79.97	72.01	21.79	114.4906	<b>108.5705</b>
Google Cloud Vision API	76.61	73.93	24.82	106.715	<b>111.8906</b>
OCR.space	78.25	63.98	22.83	105.8687	<b>107.8162</b>
Tesseract	33.22	13.10	71.33	-55.9514	<b>-41.2731</b>

## EXAMPLE CASE

Did I tell you that you are awesome  
just the way you are and that  
what I love most about you is that  
you are EVOLVING? Well... I just  
did.

### OCR.SPACE OUTPUT:

```
print(output['ParsedResults'][0]['ParsedText'])
```

Did I tell you that you are arresome  
just  
the way you are and that  
what I love most about you is that  
you are EVOLVING? kell... I just  
did.

### LLM CORRECTED OUTPUT:

```
model=prompt|llm|StrOutputParser()
print(model.invoke({'input':EXAMPLE}))
```

Did I tell you that you are awesome just the way you are and that  
what I love most about you is that  
you are evolving? Well... I just  
did.

# DOCUMENT OCR

LIAN HING STATIONERY SDN BHD  
(162761-M)  
NO.32 & 33, JALAN SR 1/9, SEKSYEN 9,  
TAMAN SERDANG RAYA,  
43300 SERI KEMBANGAN, SELANGOR  
DARUL EHSAN  
GST ID : 002139201536

## Tax Invoice

27/03/2018 No : CS-20243

F/Castell 187057-75 Tack-It  
75g- White (new) @ 5.6600

Qty	Tax	RM
2	SR	12.00

Total Amt Incl. GST @ 6% : 12.00  
Rounding Adjustment :  
Total Amt Payable : 12.00  
Paid Amount : 20.00  
Change : 8.00  
Total Qty Tender : 2

GST Summary	Amount (RM)	Tax (RM)
SR @ A	11.32	0.68
Total	11.32	0.68

THANK YOU

For any enquiry, please contact us:

MR D.T.Y. (M) SDN BHD  
(CO. REG : 860671-D)  
LOT 1051-A & 1051-B, JALAN KPB 6,  
KAWASAN PERINDUSTRIAN BALAKONG,  
43300 SERI KEMBANGAN, SELANGOR  
(TESCO PUTRA NILAI)  
-INVOICE-

KILAT AUTO ECO WASH & SHINE ES1000 1L  
WA45 /2A - 12  
9555916500133 1 X 3.11 3.11  
KILAT ECO AUTO WASH & WAX EW-1000-1L  
WA44-A - 12  
9555916500126 1 X 4.62 4.62  
WD40 277ml MQQ 2572  
WA45-A - 24  
079567600084 1 X 11.23 11.23  
KLEENSO AJAIB 99 SERAI WANYI 900G  
WD00 - 15  
9555651400385 1 X 7.45 7.45  
HANDKERCHIEF 71386#2PCS  
PI12PJ11-4 - 6/300  
9090822 1 X 4.50 4.50

Item(s) : 5 Qty(s) : 5

Total RM 30.91  
ROUNDING ADJUSTMENT -RM 0.01  
TOTAL ROUNDED RM 30.90  
CASH RM 51.00  
CHANGE RM 20.10

18-11-18 13:58 SH01 Z153 T2 R000002902  
OPERATOR TRAINEE CASHIER

EXCHANGE ARE ALLOWED WITHIN  
7 DAYS WITH RECEIPT.  
STRICTLY NO CASH REFUND.



# DOCUMENT OCR RESULTS

- **BEST PERFORMER:AWS Textract** , this highlights its excellence in extracting structured data from financial receipts and invoices,
- **Tesseract strength**: higher lev. accuracy than OCR.space in this scenario.

API	Levenshtein Accuracy(%)	Word Accuracy(%)	Character Error rate(%)	Average Score
<b>AWS Textract</b>	<b>81.98</b>	<b>44.23</b>	<b>39.66</b>	<b>73.62458</b>
Azure OCR	80.21	43.12	42.13	66.95861
Google Cloud Vision API	70.70	43.82	51.25	41.33103
OCR.space	65.74	40.25	54.83	27.37178
Tesseract	66.38	30.83	55.28	21.8764

# MULTILINGUAL OCR





# MULTILINGUAL OCR RESULTS

- **AWSTextract and Tesseract** dont support non-latin scripts , so they were excluded from the evaluation
- **BEST PERFORMER:AZURE OCR** excelled in arabic, chinese, japanese, korean recognition.
- **Google Cloud vision OCR** excelled with hindi text compared to OCR.
- **OCR.space** was the worst out of the 3, with significant struggles in japanese and chinese and not being able to recognize hindi.

Language/API	Azure word acc.	Google word acc.	OCR.space word acc.
Arabic	<b>76.60</b>	71.66	62.00
Chinese	<b>69.06</b>	63.84	45.34
Korean	<b>65.81</b>	64.47	60.12
Japanese	<b>67.00</b>	66.50	30.99
Hindi	69.01	<b>77.35</b>	2.79



# CONCLUSIONS

## Azure

- Best overall extraction quality and versatility
- Excels in structured documents, natural scene text, and handwriting
- Most comprehensive multilingual support

## AWS Textract

- Outperforms Azure for complex layouts, tables, and forms
- Limited to Latin scripts
- Requires upload to Amazon S3 (more complex setup)

## Tesseract

- Similar language limitations (struggles with non-Latin scripts)
- Poor performance on handwriting and natural scene text
- Needs extensive image preprocessing

## Google Cloud Vision OCR

- Strong language coverage and general and versatile text recognition
- Fastest model
- Slightly weaker on structured documents than AWS Textract

## OCR Space

- Cost-effective for natural scene text and handwriting
- 1MB limit on free plan, lower accuracy for structured docs
- Accuracy improves when paired with an LLM (Gemini 1.5 Flash)