INSTITUTO DE PESQUISAS TECNOLÓGICAS

COMUNICAÇÃO TÉCNICA

Nº 180139

Avanços recentes em OCR

Felipe Pacheco de Almeida Euphrásio
Adriana Camargo de Brito

Palestra apresentada no IPT, on-line. 27 slides.

Instituto de Pesquisas Tecnológicas do Estado de São Paulo S/A - IPT
Av. Prof. Almeida Prado, 532 | Cidade Universitária ou
Caixa Postal 0141 | CEP 01064-970
São Paulo | SP | Brasil | CEP 05508-901
Tel 11 3767 4374/4000 | Fax 11 3767-4099

www.ipt.br

INSTITUTO DE PESQUISAS TECNOLÓGICAS

AVANÇOS RECENTES EM OCR

Seção de Inteligência Artificial e Analytics

27.03.2026

Dr. Felipe Pacheco de Almeida Euphrásio

Colaboração: Dra. Adriana Camargo de Brito

SÃO PAULO GOVERNO DO ESTADO

QUEM SOMOS

- Felipe Pacheco de Almeida Euphrasio
  - Graduado em Engenharia de Controle e Automação (FATESF)
  - Mestre em Ciências e Tecnologias Espaciais (ITA)
  - Doutorado em Engenharia Aeronáutica e Mecânica (ITA)

- Adriana Camargo de Brito
  - Doutora em engenharia Mecânica
  - MBA em Data Science e Analytics

ROTEIRO

1. Conceitos

Definição, pipeline geral, tipos de documento e pré-processamento.

2. Modelos

OCR clássico / ML, deep learning e visual language models.

3. Evidências

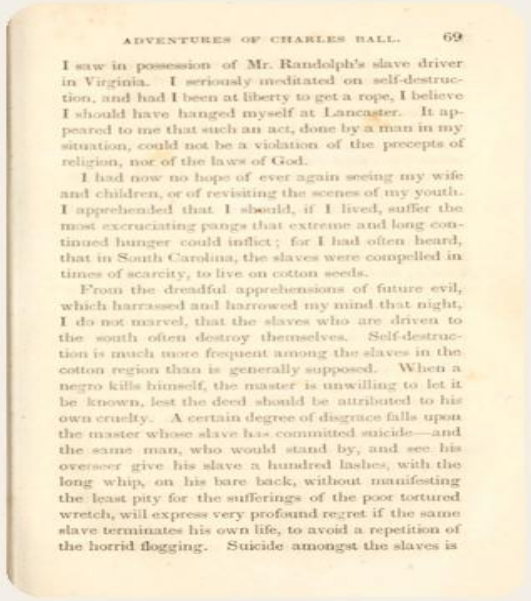
Exemplos aplicados, benchmarks, métricas e antes/depois.

4. Decisão técnica

Aplicações, limitações, riscos e arquitetura recomendada.

ROTEIRO

Pipeline OCR



ADVENTURES OF CHARLES BALL. 69

I saw in possession of Mr. Randolph's slave driver in Virginia. I seriously meditated on self-destruction, and had I been at liberty to get a rope, I believe I should have hanged myself at Lancaster. It appeared to me that such an act, done by a man in my situation, could not be a violation of the precepts of religion, nor of the laws of God.

I had now no hope of ever again seeing my wife and children, or of revisiting the scenes of my youth. I apprehended that I should, if I lived, suffer the most excruciating pangs that extreme and long continued hunger could inflict; for I had often heard, that in South Carolina, the slaves were compelled in times of scarcity, to live on cotton seeds.

From the dreadful apprehensions of future evil, which harrassed and harrowed my mind that night, I do not marvel, that the slaves who are driven to the south often destroy themselves. Self-destruction is much more frequent among the slaves in the cotton region than is generally supposed. When a negro kills himself, the master is unwilling to let it be known, lest the deed should be attributed to his own cruelty. A certain degree of disgrace falls upon the master whose slave has committed suicide—and the same man, who would stand by, and see his overseer give his slave a hundred lashes, with the long whip, on his bare back, without manifesting the least pity for the sufferings of the poor tortured wretch, will express very profound regret if the same slave terminates his own life, to avoid a repetition of the horrid flogging. Suicide amongst the slaves is

1 **Entrada**
documento ou cena

2 **Pre-processo**
contraste, ruido, deskew

3 **Deteccao**
linhas, palavras, regioes

4 **Reconhecimento**
texto bruto

5 **Pos-processo**
normalizacao e regras

ROTEIRO

OCR + VLM

OCR moderno combina leitura visual com entendimento estrutural

OCR
texto extraido

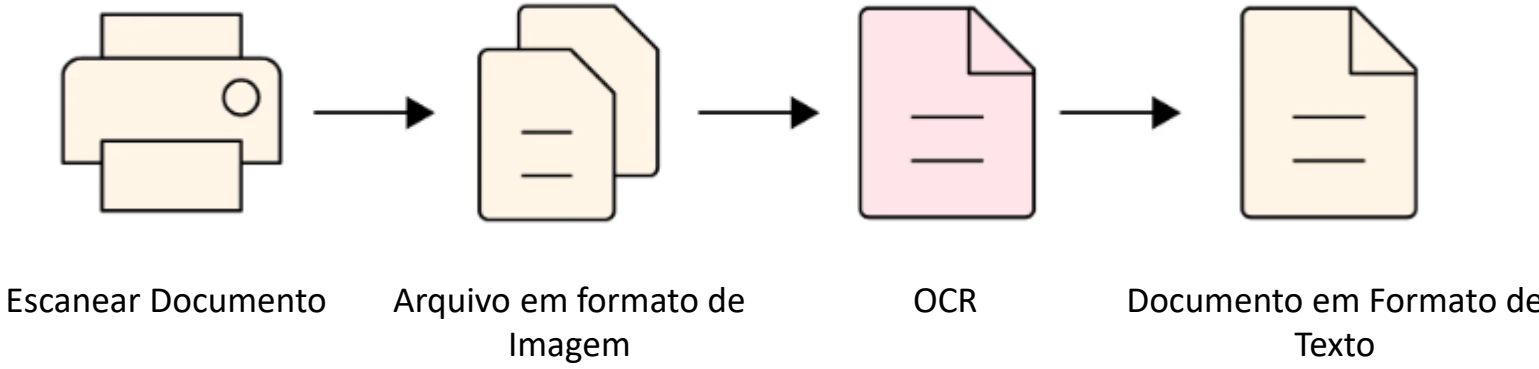
LLM
normaliza, resume, valida

Saida
JSON, tabela ou workflo

```
{
"nome": "campo do documento",
"numero": "texto validado",
"confianca": "revisar se necessario"
}
```

O QUE É OCR?

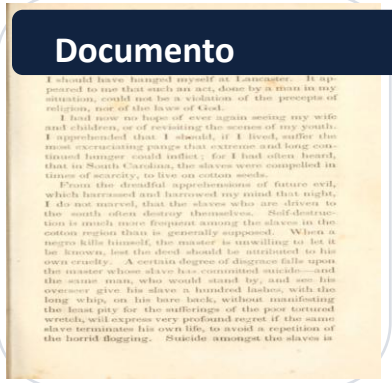
- Definição:
  - OCR (Optical Character Recognition)
  - Identifica e converte texto de imagens, fotos em texto legível por máquina
  - Permite busca, edição, extração e automação de informações.



O QUE É OCR?

- OCR converte texto visível em imagens, em saídas legíveis como campos, entidades ou JSON.
- Entrada: documento escaneado, foto de documento, cena real, manuscrito ou PDF rico.
- Meio: pré-processamento, detecção/segmentação, reconhecimento e pós-processamento.
- Saída: transcrição, campos estruturados, busca indexada ou automação documental.
- Valor real: OCR e ponte entre visão computacional, NLP e sistemas de negócio.

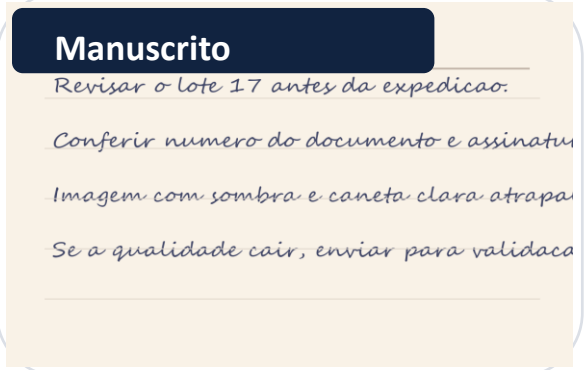
Documento



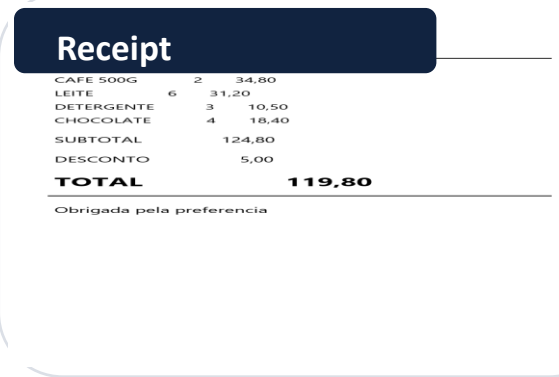
ID / Form

Cena real

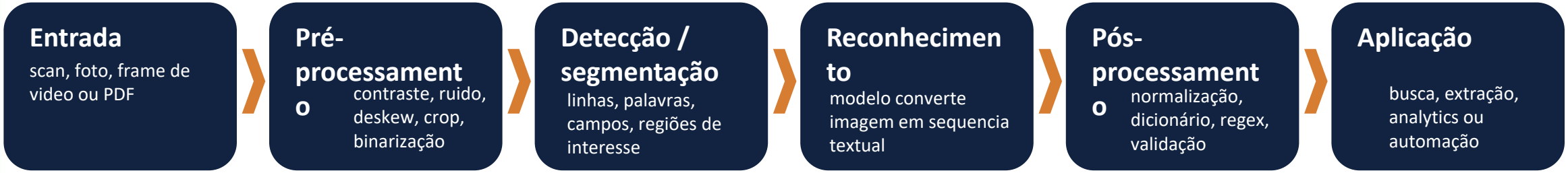
Manuscrito



Receipt



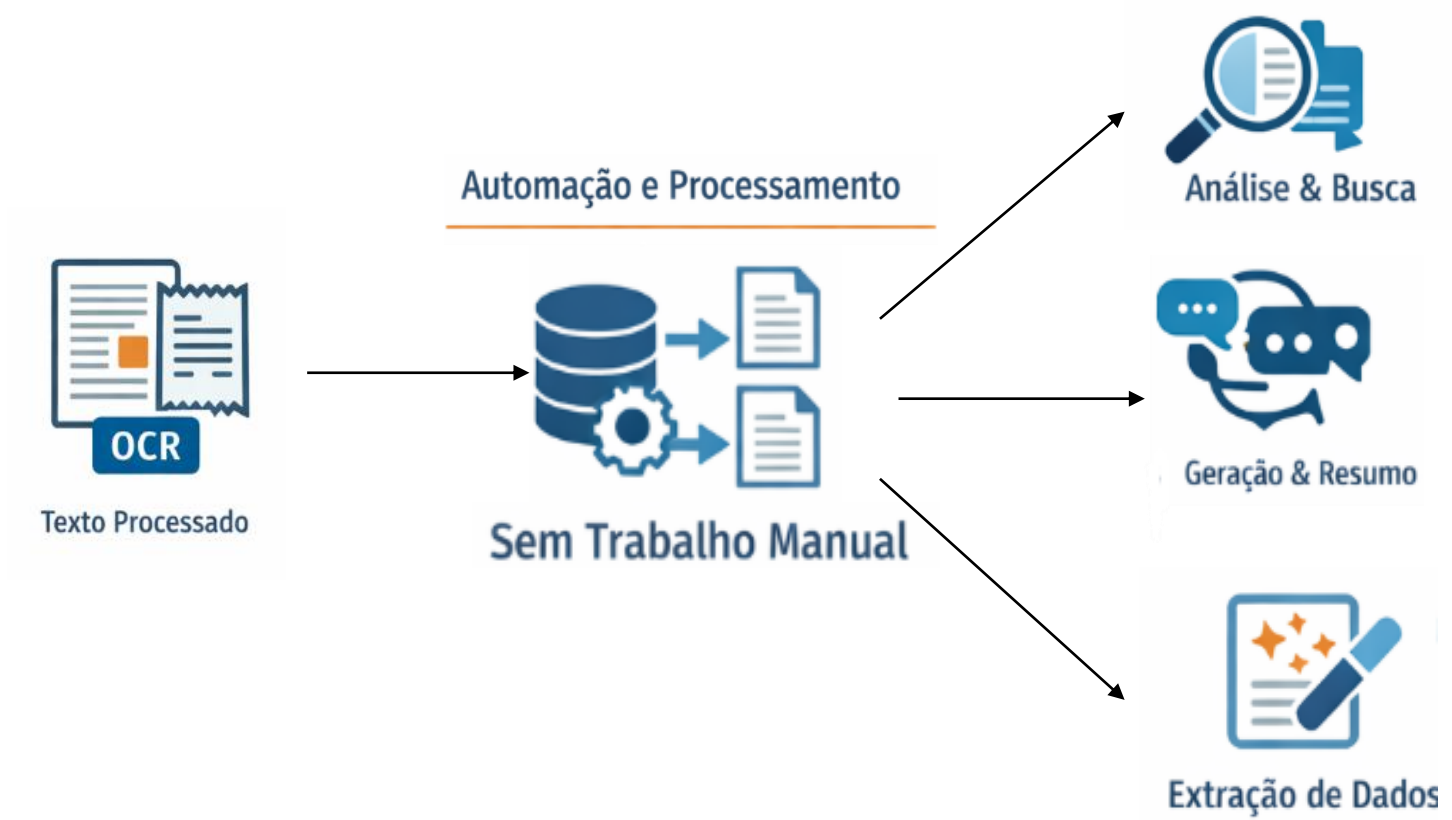
PIPELINE GERAL DE UM OCR



Etapa	Saída intermediaria	Ponto de falha tipico
Pré-processamento	imagem mais legivel	perda de detalhe, over-thresholding
Detecção	caixas / linhas / campos	texto pequeno, curvo, ocluido
Reconhecimento	texto bruto	fontes raras, blur, baixa resolução
Pós-processamento	texto confiável	regex errada, validação fraca

POR QUE OCR IMPORTA?

- Reduz trabalho manual;
- Acelera processos;
- Viabiliza automação de documentos;
- Serve de base para aplicações de IA em específico no processamento de texto usando NLP;



HISTÓRICO DO OCR

1950-1980

Templates, OCR-A/B, documentos muito padronizados e alta dependencia de regras.

1980-2010

Features manuais, segmentação forte, HMM/SVM e dicionarios especializados.

2010-2022

CNNs, CRNN, detectors de cena, transformers e ganhos fortes em robustez.

2022-2026

VLMs, OCR-free, DocVQA, extração estruturada e raciocínio multimodal.

Mudança	Antes	Depois
Unidade de previsão	caractere isolado	linha, palavra, entidade ou JSON
Fonte de conhecimento	regras e features manuais	features aprendidas e contexto multimodal
Escopo da tarefa	transcrição	transcrição + estrutura + perguntas + ação

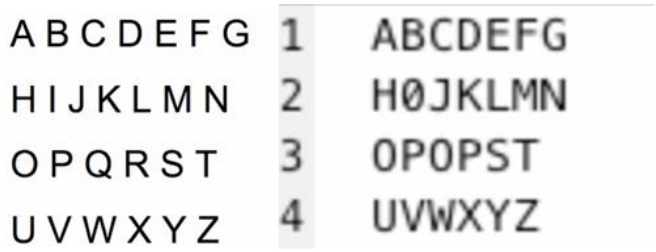
HISTÓRICO DO OCR

- OCR Clássico
- Baseado em Regras, Segmentação e Classificadores
- Antes do deep learning, os “modelos” mais comuns eram:
  - Template matching
  - Extração de características + k-NN
  - SVM
  - HMM
  - Regras heurísticas e dicionários

HISTÓRICO DO OCR

Abordagem de template matching

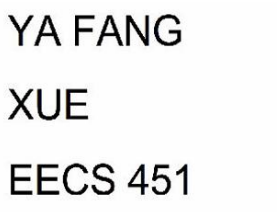
Saídas textuais



Erros nos templates

Pré Processamentos

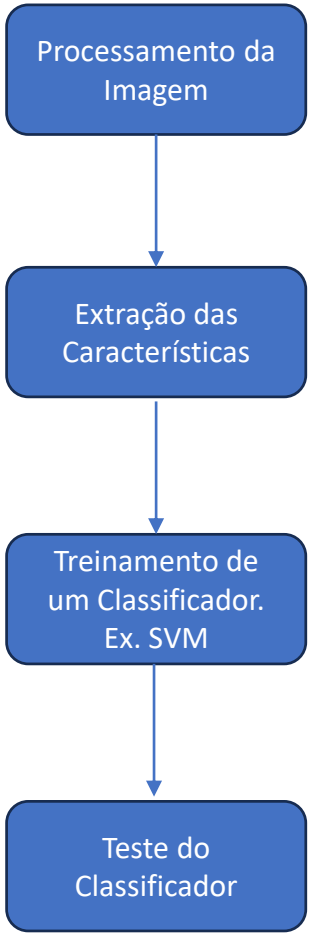
Imagem Original



Binarização

Divisão Em Linhas

Separação das Letras



[Microsoft Word - optical character recognition final report.doc]

HISTÓRICO DO OCR

Familia	Ideia central	Exemplo	Onde ainda entrega valor
Template matching	compara pixels e moldes fixos	OCR-A/B, displays, digitos	layout rigido e baixo custo
Features + classificador	zoning/HOG/projeções + kNN/SVM	placas, formulários padrão	interpretabilidade e rapidez
Modelos sequenciais	HMM/CRF + dicionário	linhas com gramática forte	reduzir ambiguidade linguistica
Motores documentais	segmentação, heuristicas e LM	Tesseract em lotes limpos	digitalização massiva de scans

Forças

Baixo custo, pouca anotação, boa explicabilidade e bom desempenho em documentos padronizados.

Limites

Muito sensível a ruido, curvatura, fonte rara, variação de layout e texto em cena real.

HISTÓRICO DO OCR

OCR baseado em Aprendizado Profundo

Camada	Modelos de referência	Papel no pipeline
Detecção	EAST, CRAFT, DBNet	localizar texto arbitrário em cenas e documentos
Reconhecimento	CRNN, SAR, TrOCR	converter imagem em sequência textual
Pipelines	EasyOCR, PaddleOCR, docTR	combinar detecção, reconhecimento e utilitários
Layout / KIE	LayoutLM, PP-Structure, parsers	usar relação espacial para campos e tabelas

Por que melhorou

Aprende features diretamente, tolera melhor rotação, escala, fontes e texto em cena real.

Custo técnico e operacional

Exige dados, GPU, tuning por cenário e benchmark contínuo por tipo de documento.

Modelos citados: EAST (Zhou et al., 2017); CRAFT (Baek et al., 2019); CRNN (Shi, Bai e Yao, 2015); TrOCR (Li et al., 2021); LayoutLM (Xu et al., 2019). Ferramentas: EasyOCR, PaddleOCR e docTR — repositórios oficiais.

HISTÓRICO DO OCR

Visual language models – OCR Multimodal

Categoria	Exemplos	O que agrega
OCR-free de documento	Donut, Nougat	imagem -> markup, JSON ou resposta sem OCR separado
LMMs gerais	familia GPT-4V / Gemini / Qwen-VL / DeepSeek	leitura + raciocínio + resposta em linguagem
Parsers multimodais	PaddleOCR-VL e doc VLMs	OCR, layout e estrutura na mesma pilha

OCRBench mostrou que VLMs são promissores, mas ainda apresentam fraquezas relevantes em texto multilíngue, manuscrito, texto não semântico e expressões matemáticas.

Forças

Entendem contexto, tabelas, relações semânticas e podem responder perguntas ou gerar JSON.

Riscos

Mais custo e latência, maior variabilidade de prompt e necessidade de governança contra hallucination.

Modelos citados: Donut (Kim et al., 2021); Nougat (Blecher et al., 2023); OCRBench (Liu et al., 2023). Ferramenta usada no exemplo prático: Together API com Llama 4 Maverick.

COMPARANDO FERRAMENTAS

Critério	Clássico / ML	Deep OCR	VLM / multimodal
Custo inicial	baixo	médio	alto
Necessidade de dados	baixa	media / alta	baixa no zero-shot, alta para dominar domínio
Tolerância a ruído	baixa	alta	media / alta com variabilidade
Entendimento de layout	baixo	médio	alto
Extração estruturada	via regra	via parser / KIE	nativa com prompt + schema
Latência e custo por página	baixo	médio	alto
Governança / explicabilidade	alta	media	baixa / media
Melhor fit	lotes limpos e padronizados	OCR produtivo em grande escala	tarefas com contexto, QA e JSON

Use clássico

Dificilmente utilizado hoje. As técnicas de deep learning sobreporam totalmente o uso clássico.

Use deep

Quando você precisa robustez, escala e texto em cena ou layout variável.

Use VLM

Quando a aplicação pede contexto, resposta e estrutura.

SÃO PAULO
GOVERNO DO ESTADO

DESAFIOS

Sombra e blur

Quebram detecção e trocam simbolos parecidos. Melhor captura e quality gate ainda são fundamentais.

Manuscrito

Ambiguidade semântica e variação de traço exigem modelo especializado, contexto e revisão.

Line items

Recibo e tabelas falham quando o OCR lê apenas texto e não entende estrutura.

Multilíngue e fonte rara

Caracteres similares e alfabetos mistos pedem treino, dicionário e benchmark específico.

Alucinação em VLM

Quando o modelo inventa campo ausente, o erro vira erro de negócio. Grounding e validação são obrigatórios.

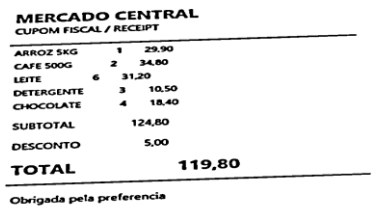
Lição central

OCR confiável depende de captura, roteamento, métricas e rota de exceção humana.

EXEMPLOS

Antes

Depois do pré-processamento



OCR bruto

```
CEN| ns
RC

3 nos E
OD.
Ecs
A
O
```

OCR após pré-processamento

```
MERCADO CENTRAL
CUPOM FISCAL / RECEIPT
mel
ARROZSKG 1 2990
carEs0OG 2 3480
LEITE 6 31,20
DETERGENTE 3 1050
CHOCOUTE 4 1840
SUBTOTAL 124,80
DESCONTO 5,00
Obrigada pela preferencia
```

Pós-processamento por regra

```
{
  "loja": "MERCADO CENTRAL",
  "subtotal": "124,80",
  "desconto": "5,00",
  "total_inferido": "119,80",
  "alerta": "TOTAL não lido; inferido
via regra subtotal - desconto"
}
```

Leitura operacional

O pré-processamento recuperou partes do subtotal e do desconto; o pós-processamento recompôs o total com rastreabilidade explícita.

EXEMPLOS

Entrada: CNH

JSON retornado pelo modelo

```
{
  "nome_completo": "NOME SOCIAL TESTE
CENTO E DEZ",
  "rg": "513584349",
  "cpf": "076.763.758-51",
  "orgao_emissor": "SSPSD"
}
```

O que o VLM agrega

Extrai campos estruturados diretamente da imagem inteira, reduzindo o esforço de parser campo a campo.

Risco principal

Mesmo com JSON limpo, CPF, datas, categoria e consistência documental ainda exigem validação determinística.

Proveniência do experimento

Modelo meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8

EXEMPLOS

Entrada: CNH

JSON retornado pelo modelo


```
{
"nome_completo":null,
"rg":"513584349",
"cpf":"07676375851",
"orgao_emissor":"SSP/SP
"}
```


O que o VLM agrega

Extrai campos estruturados diretamente da imagem inteira, reduzindo o esforço de parser campo a campo.

Risco principal

Mesmo com JSON limpo, CPF, datas, categoria e consistência documental ainda exigem validação determinística.

Proveniência do experimento

Modelo moonshotai/Kimi-K2.5

MÉTRICAS

Se tiver gabarito para comparar

- **Character Error Rate (CER)**
  - Mede a porcentagem de caracteres reconhecidos de forma incorreta
  - CER = (Inserções + deleções + substituições) / total de caracteres)
  - Ideal para documentos com muitos números, fórmulas técnicas ou códigos (onde um erro de um dígito muda tudo).

- **Word Error Rate (WER)**
  - Funciona como o CER, mas a unidade de medida é a palavra
  - Melhor para textos corridos.
  - É mais rigorosa: se uma letra de uma palavra de 10 caracteres estiver errada, a palavra inteira é contada como erro

MÉTRICAS

Se não tiver gabarito para comparar

- **F1-Score (A nível de Token)**
  - Extração de informações específicas (como nomes de ruas ou valores em tabelas). Ela equilibra:
    - **Precision:** Das palavras que o OCR leu, quantas estão certas?
    - **Recall:** De todas as palavras que estavam no papel, quantas o OCR conseguiu ler?

- **LM-as-a-Judge (O "Juiz" IA)**
- Você passa o texto extraído pelo OCR para um modelo robusto (como o GPT-4o) e pede para ele avaliar.
  - **Prompt:** "Avalie de 0 a 10 a coerência deste texto extraído de um PDF. Identifique palavras truncadas, caracteres especiais fora de contexto ou quebras de linha erradas."
  - **Vantagem:** Consegue identificar erros de caracteres

CONCLUSÕES

- Não existe um único OCR vencedor. A pilha correta depende do documento, da saída desejada e do custo de errar.

Documentos limpos

Pipelines baratos ainda resolvem muito quando layout e captura são estáveis.

Contexto e estrutura

VLMs destravam QA, tabelas, extração em JSON e entendimento de documento.

Cenas e layouts

Deep OCR virou o padrão operacional para robustez, escala e variabilidade.

Futuro

Pilhas híbridas, menores e governadas por confidence, benchmark e human-in-the-loop.

REFERÊNCIAS

SMITH, Ray. An overview of the Tesseract OCR engine. In: INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, 9., 2007, Curitiba. Proceedings [...]. Los Alamitos: IEEE Computer Society, 2007. p. 629-633. DOI: 10.1109/ICDAR.2007.4376991.

SHI, Baoguang; BAI, Xiang; YAO, Cong. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. arXiv, 2015. Disponível em: https://arxiv.org/abs/1507.05717. Acesso em: 25 mar. 2026.

ZHOU, Xinyu et al. EAST: An efficient and accurate scene text detector. arXiv, 2017. Disponível em: https://arxiv.org/abs/1704.03155. Acesso em: 25 mar. 2026.

BAEK, Youngmin et al. Character region awareness for text detection. arXiv, 2019. Disponível em: https://arxiv.org/abs/1904.01941. Acesso em: 25 mar. 2026.

LIAO, Minghui et al. Real-time scene text detection with differentiable binarization. arXiv, 2019. Disponível em: https://arxiv.org/abs/1911.08947. Acesso em: 25 mar. 2026.

REFERÊNCIAS

JAUME, Guillaume; EKENEL, Hazim Kemal; THIRAN, Jean-Philippe. FUNSD: A dataset for form understanding in noisy scanned documents. arXiv, 2019. Disponível em: https://arxiv.org/abs/1905.13538. Acesso em: 25 mar. 2026.

MATHEW, Minesh; KARATZAS, Dimosthenis; JAWAHAR, C. V. DocVQA: a dataset for VQA on document images. In: WACV, 2021. Disponível em: https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html. Acesso em: 25 mar. 2026.

LI, Minghao et al. TrOCR: Transformer-based optical character recognition with pre-trained models. arXiv, 2021. Disponível em: https://arxiv.org/abs/2109.10282. Acesso em: 25 mar. 2026.

KIM, Geewook et al. OCR-free document understanding transformer. arXiv, 2021. Disponível em: https://arxiv.org/abs/2111.15664. Acesso em: 25 mar. 2026.

BLECHER, Lukas et al. Nougat: Neural optical understanding for academic documents. arXiv, 2023. Disponível em: https://arxiv.org/abs/2308.13418. Acesso em: 25 mar. 2026.

LIU, Yuliang et al. OCRBench: On the hidden mystery of OCR in large multimodal models. arXiv, 2023. Disponível em: https://arxiv.org/abs/2305.07895. Acesso em: 25 mar. 2026.

REFERÊNCIAS

XU, Yiheng et al. LayoutLM: pre-training of text and layout for document image understanding. arXiv, 2019. Disponível em: https://arxiv.org/abs/1912.13318. Acesso em: 25 mar. 2026.

JAIDEDAI. EasyOCR. GitHub, 2026. Disponível em: https://github.com/JaidedAI/EasyOCR. Acesso em: 25 mar. 2026.

PADDLEPADDLE. PaddleOCR. GitHub, 2026. Disponível em: https://github.com/PaddlePaddle/PaddleOCR. Acesso em: 25 mar. 2026.

MINDEE. docTR. GitHub, 2026. Disponível em: https://github.com/mindee/doctr. Acesso em: 25 mar. 2026

Obrigado!

- fpacheco@ipt.br
- adrianab@ipt.br

linkedin.com/school/iptsp/

instagram.com/ipt_oficial/

youtube.com/@IPTbr/

www.ipt.br

www.linkedin.com/in/felipeeuphrasio

https://github.com/felipephelp/Extractor_Tools

ipt INSTITUTO DE PESQUISAS TECNOLÓGICAS

SÃO PAULO GOVERNO DO ESTADO