

Universidade Federal do ABC  
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas  
Trabalho de Graduação em Engenharia de Informação

**Estudo e aplicação de redes neurais profundas  
na solução de detecção e reconhecimento de  
texto em cenas**

**Matheus Milani**

**Maio de 2022  
Santo André - SP**



Matheus Milani

**Estudo e aplicação de redes neurais profundas na solução  
de detecção e reconhecimento de texto em cenas**

**Trabalho de Graduação** apresentado para conclusão da Graduação em Engenharia de Informação, como parte dos requisitos necessários para a obtenção do Título de Bacharel em Engenharia de Informação.

Universidade Federal do ABC  
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas  
Trabalho de Graduação em Engenharia de Informação

Orientador: Murilo Bellezoni Loiola

Santo André - SP  
Maio de 2022

# Resumo

Segundo a ABNT, o resumo deve ressaltar o objetivo, o método, os resultados e as conclusões do documento. A ordem e a extensão destes itens dependem do tipo de resumo (informativo ou indicativo) e do tratamento que cada item recebe no documento original. O resumo deve ser precedido da referência do documento, com exceção do resumo inserido no próprio documento. Umas 10 linhas (...) As palavras-chave devem figurar logo abaixo do resumo, antecedidas da expressão Palavras-chave:, separadas entre si por ponto e finalizadas também por ponto.

**Palavras-chaves:** latex. abntex. editoração de texto.

# Abstract

This is the english abstract.

**Keywords:** latex, abntex, text editoration.



# Listas de ilustrações

Figura 1 – Ilustração das etapas de geração dos arquivos de ground truth para a etapa de treino do CRAFT. Fonte [1] . . . . .	8
Figura 2 – Exemplificação do passo a passo para geração de anotações a nível de caracteres durante a etapa de treino do CRAFT. Fonte [1]. . . . .	9
Figura 3 – Ilustração do pipeline de reconhecimento do CRNN. Fonte [2]. . . . .	10
Figura 4 – Exemplo de resultado de reconhecimento. Imagem 52 do conjunto de validação. . . . .	16
Figura 5 – Comparação entre entrada e saída sobre a imagem 5 do set de validação do ICDAR 2011 . . . . .	17
Figura 6 – Exemplo de imagem com instâncias de texto bem pequenas em resolução. Imagem 28 do ICDAR 2011 . . . . .	18
Figura 7 – Demonstração de um falso positivo durante a avaliação da solução contra o dataset ICDAR 2013 . . . . .	19
Figura 8 – Exemplo de imagem com artefatos que trouxeram dificuldades para a localização correta do texto. . . . .	20
Figura 9 – Exemplo de texto sob vidro, com plano de fundo desafiadores para o reconhecimento. . . . .	20
Figura 10 – Exemplo de imagem com texto em desfoco . . . . .	20
Figura 11 – Exemplo de imagem autoral onde a solução apresentou dificuldades com texto curvo e estilizado. Textos reconhecidos: “sney” e “intmalakingon”. . . . .	21
Figura 12 – Exemplo de imagem autoral onde uma região de texto muito longa não foi extraída muito bem, principalmente no início da palavra, o que dificultou o reconhecimento. Texto reconhecido: “permmusem”. . . . .	21
Figura 13 – Exemplo de imagem autoral com fontes bastante estilizadas que trouxeram dificuldade tanto para a detecção, quanto para o reconhecimento. Textos reconhecidos: “sey” e “fpan”. . . . .	22
Figura 14 – Exemplo de reconhecimento com sucesso em condições desafiadoras em imagem autoral. Textos reconhecidos: “disneys”, “electrical”e “parade”. . . . .	22
Figura 15 – Exemplo de imagem autoral com alta precisão de reconhecimento. Textos reconhecidos: “nelson”, “rolihlahla” “mandela”, “1918”, “2013”, “hamba”, “kahle”, “madiba”, “ve”, “honour”, “your”, “legacy”, “apartheid”, “museum”. . . . .	22



# **Lista de tabelas**

Tabela 1 – Avaliação de resultados sobre a base ICDAR 2011.	16
Tabela 2 – Avaliação de resultados sobre a base ICDAR 2013.	19



# **Lista de abreviaturas e siglas**

ABNT      Associação Brasileira de Normas Técnicas

abnTeX      Normas para TeX



# List of symbols

$\Gamma$  Greek letter Gamma

$\Lambda$  Lambda

$\zeta$  Greek letter minuscule zeta

$\in$  Pertains



# Sumário

<b>1</b>	<b>INTRODUÇÃO</b>	<b>1</b>
<b>1.1</b>	<b>Scene Text Recognition</b>	<b>2</b>
<b>1.2</b>	<b>Objetivo</b>	<b>2</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>5</b>
<b>2.1</b>	<b>Aprendizado de Máquina e <i>Deep Learning</i></b>	<b>5</b>
2.1.1	Linguagens de Programação e <i>Frameworks</i>	6
<b>2.2</b>	<b>Evolução do <i>Scene Text Recognition</i></b>	<b>6</b>
2.2.1	Detecção de Texto	6
2.2.1.1	CRAFT	7
2.2.2	Reconhecimento de Texto	9
2.2.2.1	CRNN	9
<b>2.3</b>	<b>Considerações Finais</b>	<b>11</b>
<b>3</b>	<b>METODOLOGIA</b>	<b>13</b>
<b>3.1</b>	<b>Considerações Finais</b>	<b>13</b>
<b>4</b>	<b>RESULTADOS E DISCUSSÃO</b>	<b>15</b>
<b>4.1</b>	<b>ICDAR 2011</b>	<b>15</b>
4.1.1	ICDAR 2013	18
4.1.2	Imagens autorais	20
<b>5</b>	<b>CONCLUSÃO</b>	<b>23</b>
	<b>REFERÊNCIAS</b>	<b>25</b>



# 1 Introdução

Termos como inteligência artificial e aprendizado de máquina já são praticamente constantes do cotidiano humano, mesmo que muitas vezes não visível, já estão presentes em muitas aplicações que muitas vezes sequer imaginamos, desde a assistente virtual dos dispositivos móveis e eletronicos, até sistemas capazes de embasar diagnósticos médicos, como por exemplo uma aplicação capaz de auxiliar em diagnósticos de COVID-19 durante a pandemia do vírus SARS-COV2 [3].

Um escopo de aplicação de técnicas de aprendizado de máquina que tem evoluído bastante com o crescimento da popularidade e acessibilidade dos conceitos que envolvem *machine learning* são aplicações voltadas para reconhecimento de texto.

A escrita com certeza foi uma das grandes habilidades que a humanidade desenvolveu que mudou como relações e sociedades funcionavam, sendo adotada para transmissão e armazenagem de dados e informação, meio de comunicação e expressão.

Com os avanços da tecnologia e em especial, dos computadores, um grande desafio emergiu: Como fazer com que computadores entendam o que está escrito em documentos físicos? Umas das primeiras patentes para soluções de OCR (abreviação de *Optical Character Recognition*) data de 1929 [4], mas isso não nega o fato que a capacidade de transportar texto do meio físico para o meio digital de forma eficiente é um desafio interessante e que motiva pesquisas até hoje.

Em linhas gerais, o reconhecimento óptico de caracteres é uma ampla tarefa de reconhecimento de padrões e, para que máquinas consigam identificar os padrões presentes na instâncias de texto, elas precisam conhecer ao menos algumas características dos caracteres e do texto que serão reconhecidos. Para exemplificar, a Reading Machine de Tauschek [4] era uma aparelho mecânico que utilizava um disco de comparação, que continha o gabarito de cada um dos caracteres do alfabeto suportado. Esse foi o meio de "ensinar" a máquina a reconhecer um dado caracter.

Muitos anos depois, soluções ainda tem a missão de "treinar" computadores a identificar as características de caracteres e de textos como um todo e a principal ferramenta utilizada nos dias atuais são métodos sob o domínio de aprendizado de máquina, justamente pela capacidade de predição desses algoritmos dado um processo de treinamento. Ao longo deste trabalho de graduação outros conceitos que circundam o topoico de aprendizado de maquina serão introduzidos com um pouco mais de profundidade.

## 1.1 Scene Text Recognition

Um sub-conjunto de casos do espaço de aplicações de reconhecimento óptico de caracteres ganhou bastante tração no última década, impulsionada pelo alto poder computacional dos dispositivos modernos, acelerados por unidades gráficas, e a acessibilidade ao desenvolvimento de soluções de aprendizado de máquina. Esse sub-conjunto é conhecido como STR (abreviação de *Scene Text Recognition*, em inglês). Uma analogia para o STR seria aplicar soluções de OCR diretamente de fotos capturadas por uma câmera de um dispositivo móvel.

Como o nome sugere, STR classifica no sub-conjunto onde o problema a ser resolvido é a detecção e o reconhecimento do texto em imagens cotidianas, em cenas. A diferença entre um problema de STR comparado ao caso mais comum de OCR é, em termos simples, a aparência do texto e como ele será observado. Em uma imagem de cena, como por exemplo uma imagem da faixada de um supermercado, podemos ter textos em diferentes tamanhos, com diversas fontes, cores e orientações. Adicionalmente, por estarem muitas vezes sob influência do ambiente onde estão inseridos, outros fatores influenciam a observação desse texto, como iluminação, oclusão, danos devido ao clima, etc.

As soluções para problemas de STR, para lidarem com o nível de generalização necessário para reconhecer texto nos mais diversos casos, são largamente baseadas nos conceitos de deep learning, que demonstram ser capazes de ir um passo à frente no quesito reconhecimento de padrões em comparação às técnicas clássicas de processamento de imagem e conseguirem ser aplicadas com eficiência sobre imagens, sendo a aplicação das redes neurais convolucionais, comumente abreviadas para CNN (*Convolutional Neural Networks*, em inglês), um divisor de águas na evolução dessas soluções. [5, 6].

Assim, dada a importância do deep-learning, e em especial das CNNs nesse contexto de detecção e reconhecimento de texto, a próxima seção irá apresentar alguns conceitos básicos associados a essas estruturas.

## 1.2 Objetivo

Nulla malesuada risus ut urna. Aenean pretium velit sit amet metus. Duis iaculis. In hac habitasse platea dictumst. Nullam molestie turpis eget nisl. Duis a massa id pede dapibus ultricies. Sed eu leo. In at mauris sit amet tortor bibendum varius. Phasellus justo risus, posuere in, sagittis ac, varius vel, tortor. Quisque id enim. Phasellus consequat, libero pretium nonummy fringilla, tortor lacus vestibulum nunc, ut rhoncus ligula neque id justo. Nullam accumsan euismod nunc. Proin vitae ipsum ac metus dictum tempus. Nam ut wisi. Quisque tortor felis, interdum ac, sodales a, semper a, sem. Curabitur in velit sit amet dui tristique sodales. Vivamus mauris pede, lacinia eget, pellentesque quis,

scelerisque eu, est. Aliquam risus. Quisque bibendum pede eu dolor.



## 2 Fundamentação Teórica

Maecenas accumsan dapibus sapien. Duis pretium iaculis arcu. Curabitur ut lacus. Aliquam vulputate. Suspendisse ut purus sed sem tempor rhoncus. Ut quam dui, fringilla at, dictum eget, ultricies quis, quam. Etiam sem est, pharetra non, vulputate in, pretium at, ipsum. Nunc semper sagittis orci. Sed scelerisque suscipit diam. Ut volutpat, dolor at ullamcorper tristique, eros purus mollis quam, sit amet ornare ante nunc et enim.

Phasellus fringilla, metus id feugiat consectetur, lacus wisi ultrices tellus, quis lobortis nibh lorem quis tortor. Donec egestas ornare nulla. Mauris mi tellus, porta faucibus, dictum vel, nonummy in, est. Aliquam erat volutpat. In tellus magna, porttitor lacinia, molestie vitae, pellentesque eu, justo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Sed orci nibh, scelerisque sit amet, suscipit sed, placerat vel, diam. Vestibulum nonummy vulputate orci. Donec et velit ac arcu interdum semper. Morbi pede orci, cursus ac, elementum non, vehicula ut, lacus. Cras volutpat. Nam vel wisi quis libero venenatis placerat. Aenean sed odio. Quisque posuere purus ac orci. Vivamus odio. Vivamus varius, nulla sit amet semper viverra, odio mauris consequat lacus, at vestibulum neque arcu eu tortor. Donec iaculis tincidunt tellus. Aliquam erat volutpat. Curabitur magna lorem, dignissim volutpat, viverra et, adipiscing nec, dolor. Praesent lacus mauris, dapibus vitae, sollicitudin sit amet, nonummy eget, ligula.

Cras egestas ipsum a nisl. Vivamus varius dolor ut dolor. Fusce vel enim. Pellentesque accumsan ligula et eros. Cras id lacus non tortor facilisis facilisis. Etiam nisl elit, cursus sed, fringilla in, congue nec, urna. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer at turpis. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Duis fringilla, ligula sed porta fringilla, ligula wisi commodo felis, ut adipiscing felis dui in enim. Suspendisse malesuada ultrices ante. Pellentesque scelerisque augue sit amet urna. Nulla volutpat aliquet tortor. Cras aliquam, tellus at aliquet pellentesque, justo sapien commodo leo, id rhoncus sapien quam at erat. Nulla commodo, wisi eget sollicitudin pretium, orci orci aliquam orci, ut cursus turpis justo et lacus. Nulla vel tortor. Quisque erat elit, viverra sit amet, sagittis eget, porta sit amet, lacus.

### 2.1 Aprendizado de Máquina e *Deep Learning*

Cras egestas ipsum a nisl. Vivamus varius dolor ut dolor. Fusce vel enim. Pellentesque accumsan ligula et eros. Cras id lacus non tortor facilisis facilisis. Etiam nisl elit, cursus sed, fringilla in, congue nec, urna. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer at turpis. Cum sociis natoque penatibus

et magnis dis parturient montes, nascetur ridiculus mus. Duis fringilla, ligula sed porta fringilla, ligula wisi commodo felis, ut adipiscing felis dui in enim. Suspendisse malesuada ultrices ante. Pellentesque scelerisque augue sit amet urna. Nulla volutpat aliquet tortor. Cras aliquam, tellus at aliquet pellentesque, justo sapien commodo leo, id rhoncus sapien quam at erat. Nulla commodo, wisi eget sollicitudin pretium, orci orci aliquam orci, ut cursus turpis justo et lacus. Nulla vel tortor. Quisque erat elit, viverra sit amet, sagittis eget, porta sit amet, lacus.

### 2.1.1 Linguagens de Programação e *Frameworks*

Cras egestas ipsum a nisl. Vivamus varius dolor ut dolor. Fusce vel enim. Pellentesque accumsan ligula et eros. Cras id lacus non tortor facilisis facilisis. Etiam nisl elit, cursus sed, fringilla in, congue nec, urna. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer at turpis. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Duis fringilla, ligula sed porta fringilla, ligula wisi commodo felis, ut adipiscing felis dui in enim. Suspendisse malesuada ultrices ante. Pellentesque scelerisque augue sit amet urna. Nulla volutpat aliquet tortor. Cras aliquam, tellus at aliquet pellentesque, justo sapien commodo leo, id rhoncus sapien quam at erat. Nulla commodo, wisi eget sollicitudin pretium, orci orci aliquam orci, ut cursus turpis justo et lacus. Nulla vel tortor. Quisque erat elit, viverra sit amet, sagittis eget, porta sit amet, lacus.

## 2.2 Evolução do *Scene Text Recognition*

Cras egestas ipsum a nisl. Vivamus varius dolor ut dolor. Fusce vel enim. Pellentesque accumsan ligula et eros. Cras id lacus non tortor facilisis facilisis. Etiam nisl elit, cursus sed, fringilla in, congue nec, urna. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer at turpis. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Duis fringilla, ligula sed porta fringilla, ligula wisi commodo felis, ut adipiscing felis dui in enim. Suspendisse malesuada ultrices ante. Pellentesque scelerisque augue sit amet urna. Nulla volutpat aliquet tortor. Cras aliquam, tellus at aliquet pellentesque, justo sapien commodo leo, id rhoncus sapien quam at erat. Nulla commodo, wisi eget sollicitudin pretium, orci orci aliquam orci, ut cursus turpis justo et lacus. Nulla vel tortor. Quisque erat elit, viverra sit amet, sagittis eget, porta sit amet, lacus.

### 2.2.1 Detecção de Texto

Vivamus eu tellus sed tellus consequat suscipit. Nam orci orci, malesuada id, gravida nec, ultricies vitae, erat. Donec risus turpis, luctus sit amet, interdum quis, porta sed,

ipsum. Suspendisse condimentum, tortor at egestas posuere, neque metus tempor orci, et tincidunt urna nunc a purus. Sed facilisis blandit tellus. Nunc risus sem, suscipit nec, eleifend quis, cursus quis, libero. Curabitur et dolor. Sed vitae sem. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Maecenas ante. Duis ullamcorper enim. Donec tristique enim eu leo. Nullam molestie elit eu dolor. Nullam bibendum, turpis vitae tristique gravida, quam sapien tempor lectus, quis pretium tellus purus ac quam. Nulla facilisi.

### 2.2.1.1 CRAFT

Character Region Awareness for Text Detection [1], ou simplesmente CRAFT, é um método de detecção publicado por Youngmin Baek et al. (2019), integrantes do time de pesquisa da empresa coreana Naver Corporation<sup>[7]</sup>, que apresentou resultados bastante competitivos quando comparado aos resultados estado-da-arte do momento, superando as melhores soluções do momento em acurácia de detecção com desempenho, capacidade de detecção em frames por segundo, competitiva com os melhores métodos já publicados.

Youngmin Baek et al. introduz um método de detecção a nível de caractere onde um modelo de rede convolucional FCN é criado a partir da reconhecida rede de extração de feature VGG-16<sup>[8]</sup>. A arquitetura da rede CRAFT se inspirou na rede U-Net<sup>[9]</sup> ao introduzir skip-connections, agregando características de alto e baixo nível entre os blocos de upsampling, que decodificam o mapa de predição em dois resultados ao final da rede:

- Mapa de predição de região de carácter (*Character Region Score*): probabilidades de cada pixel está localizado no centro de um carácter
- Mapa de predição de afinidade entre caracteres (*Character Afinity Score*): Probabilidades de cada pixel está localizado no centro da região entre caracteres

Como o resultado da rede é bem específico, as imagens de *ground truth* são geradas a partir de processamento de imagens. A partir das *bounding boxes* de cada caractere, um *heatmap* de uma distribuição gaussiana é projetada dentro de cada região de caractere, representando uma distribuição de probabilidade onde o centro da distribuição é o centro da região de caractere. Com isso tem-se o gabarito do primeiro resultado da rede. O *ground truth* para as regiões de afinidade envolve novamente projetar um heatmap gaussiano em uma região que, agora, é calculada em tempo de execução a partir dos *bounding boxes* de cada caractere. A Fig. 1 ilustra o método utilizado, que se baseia em calcular uma região retangular entre caracteres utilizando os centróides dos caracteres vizinhos e centróides de triângulos gerados em cada caractere.

Para treinar a rede, os autores se utilizaram de uma estratégia de aprendizado levemente supervisionado. Como as principais bases de imagens para treinamento não

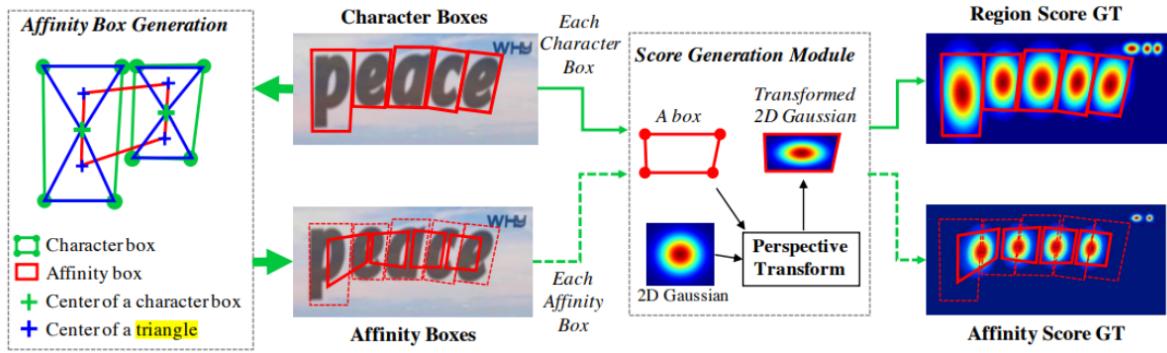


Figura 1 – Ilustração das etapas de geração dos arquivos de ground truth para a etapa de treinamento do CRAFT. Fonte [1]

contam com anotações a nível de caracteres, a rede primariamente é treinada com imagens com texto sintético, usando o dataset SynthText [10].

Para refinar o treino em datasets com imagens de cenas reais, os autores utilizam a própria rede treinada em texto sintético para predizer as regiões de caracteres das imagens de cena para gerar anotações a nível de caracteres para as imagens dos datasets utilizados com auxílio de métodos de processamento de imagem, conforme exemplificado na Fig. 2. O processo contém as seguintes etapas:

- *Cropping:* Extração das palavras que possuem região descrita nos arquivos de *ground truth* dos datasets de benchmark.
- *Character Split:* Processo de localização e segregação de cada caractere detectado pela rede treinada em base de dados sintética. A rede a partir das imagens provenientes da etapa de *Cropping*, predizendo as regiões onde a probabilidade de existir um caractere. Com a localização dessas regiões, é aplicado o algoritmo de segmentação conhecido como Watershed [11], cujo objetivo é expandir a área de um caractere a partir do centro da região de maior probabilidade até que as áreas de caracteres adjacentes se encontrem. Isso faz com que seja possível ajustar um bounding-box em volta de cada caractere observado.
- *Unwarping:* Uma vez em posse da capacidade de localizar todos os caracteres, obtida através da etapa de *Character Split*, pode-se projetar as coordenadas para as *bounding-boxes* de cada caractere de volta para a imagem original aplicando as operações inversas às aplicadas na etapa de *Cropping*.

Com essas labels geradas sobre as imagens reais dos datasets de benchmark, os gabaritos para o treinamento do modelo completo são gerados conforme explicado anteriormente e o treinamento da rede é refinado com esses novos exemplos.

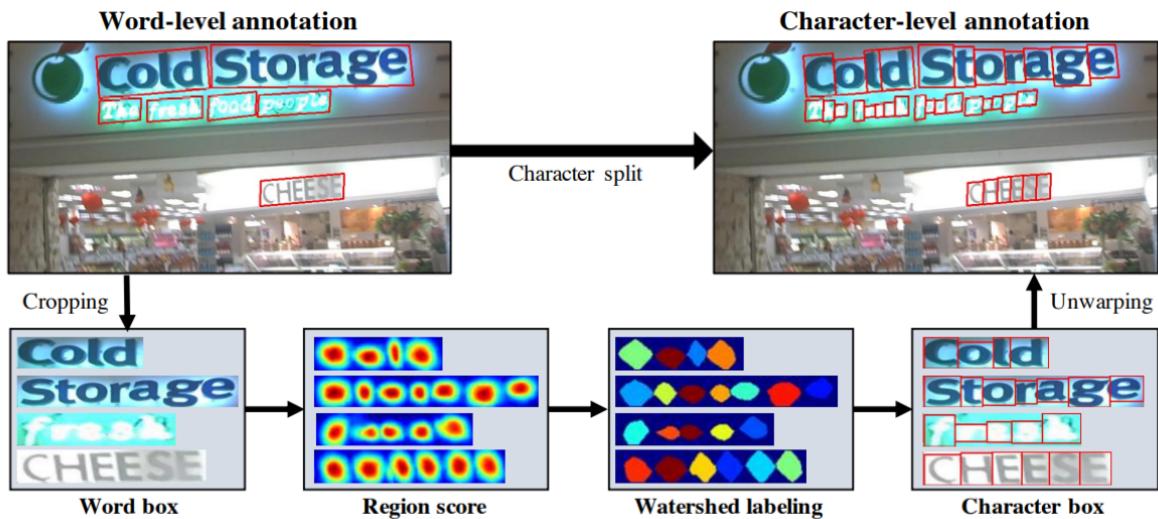


Figura 2 – Exemplificação do passo a passo para geração de anotações a nível de caracteres durante a etapa de treino do CRAFT. Fonte [1].

O CRAFT conta com um pós-processamento bastante simplificado em cima dos mapas de probabilidade que são gerados pela rede com o intuito de calcular os bounding boxes do texto localizado, que envolve, novamente com auxílio de métodos de visão computacional e processamento de imagem. Usando binarização e categorização, é possível unir as regiões de caracteres e de afinidade para extrair as coordenadas dos menores retângulos que encapsulam o resultado dessa união.

### 2.2.2 Reconhecimento de Texto

Vivamus eu tellus sed tellus consequat suscipit. Nam orci orci, malesuada id, gravida nec, ultricies vitae, erat. Donec risus turpis, luctus sit amet, interdum quis, porta sed, ipsum. Suspendisse condimentum, tortor at egestas posuere, neque metus tempor orci, et tincidunt urna nunc a purus. Sed facilisis blandit tellus. Nunc risus sem, suscipit nec, eleifend quis, cursus quis, libero. Curabitur et dolor. Sed vitae sem. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Maecenas ante. Duis ullamcorper enim. Donec tristique enim eu leo. Nullam molestie elit eu dolor. Nullam bibendum, turpis vitae tristique gravida, quam sapien tempor lectus, quis pretium tellus purus ac quam. Nulla facilisi.

#### 2.2.2.1 CRNN

Convolutional Recurrent Neural Networks [2], introduzido por Baoguang Shi et al. é uma solução para o problema de reconhecimento de texto bastante popular, sendo bastante citada em novos trabalhos e sempre presente em trabalhos comparativos. Este método veio para resolver grandes dificuldade das soluções anteriores, por exemplo: lidar com entradas e saídas de comprimentos variados, possibilitar o aprendizado conhecido

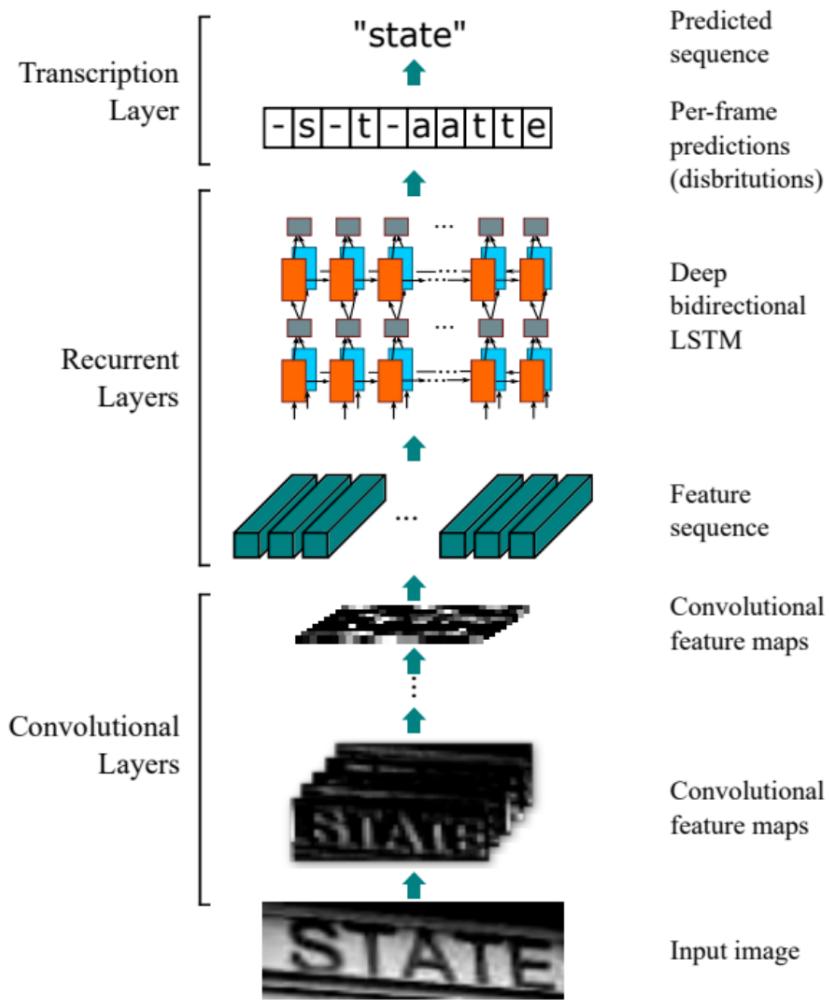


Figura 3 – Ilustração do pipeline de reconhecimento do CRNN. Fonte [2].

como end-to-end, isto é, aplicar uma função de perda sobre o resultado do reconhecimento e essa perda ser propagada para o aprendizado da rede extratora de características.

O CRNN, como o nome sugere, une uma rede neural convolucional, sem as camadas de predição totalmente conectadas, para gerar os feature maps sobre a imagem de entrada. Esses mapas são sequenciados para que alimentem a rede neural recorrente (RNN), que é responsável por conseguir decodificar cada sequência em um possível caractere. A Fig. 3 ilustra o pipeline de processamento que essa arquitetura executa.

A rede RNN que é implementada no CRNN faz uso de atributos LSTM (*Long-Short Term Memory*) bi-direcional, pois por estar visando reconhecimento de texto em cenas, muitas vezes carregar o contexto de sequências passadas e futuras é importante para diferenciar caracteres ou possibilitar o reconhecimento de um caractere, por exemplo, uma letra um pouco mais larga.

Os resultados obtidos foram bastante competitivos quando comparados aos métodos de reconhecimento anteriores, até mesmo superiores aos que já faziam uso dos conceitos

de deep learning, com o adicional de prover meios de aprendizado integrado da rede convolucional e recorrente como uma unidade, além de um modelo bem mais enxuto em número de parâmetros<sup>[2]</sup>.

Por lidar com o reconhecimento como um problema de sequência, o modelo pressupõe que a orientação do texto é necessariamente da esquerda para a direita, isso leva a uma limitação, que seria reconhecer textos com não exatamente horizontais e retilíneos.

## 2.3 Considerações Finais

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.



## 3 Metodologia

Maecenas accumsan dapibus sapien. Duis pretium iaculis arcu. Curabitur ut lacus. Aliquam vulputate. Suspendisse ut purus sed sem tempor rhoncus. Ut quam dui, fringilla at, dictum eget, ultricies quis, quam. Etiam sem est, pharetra non, vulputate in, pretium at, ipsum. Nunc semper sagittis orci. Sed scelerisque suscipit diam. Ut volutpat, dolor at ullamcorper tristique, eros purus mollis quam, sit amet ornare ante nunc et enim.

Phasellus fringilla, metus id feugiat consectetur, lacus wisi ultrices tellus, quis lobortis nibh lorem quis tortor. Donec egestas ornare nulla. Mauris mi tellus, porta faucibus, dictum vel, nonummy in, est. Aliquam erat volutpat. In tellus magna, porttitor lacinia, molestie vitae, pellentesque eu, justo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Sed orci nibh, scelerisque sit amet, suscipit sed, placerat vel, diam. Vestibulum nonummy vulputate orci. Donec et velit ac arcu interdum semper. Morbi pede orci, cursus ac, elementum non, vehicula ut, lacus. Cras volutpat. Nam vel wisi quis libero venenatis placerat. Aenean sed odio. Quisque posuere purus ac orci. Vivamus odio. Vivamus varius, nulla sit amet semper viverra, odio mauris consequat lacus, at vestibulum neque arcu eu tortor. Donec iaculis tincidunt tellus. Aliquam erat volutpat. Curabitur magna lorem, dignissim volutpat, viverra et, adipiscing nec, dolor. Praesent lacus mauris, dapibus vitae, sollicitudin sit amet, nonummy eget, ligula.

Cras egestas ipsum a nisl. Vivamus varius dolor ut dolor. Fusce vel enim. Pellentesque accumsan ligula et eros. Cras id lacus non tortor facilisis facilisis. Etiam nisl elit, cursus sed, fringilla in, congue nec, urna. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer at turpis. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Duis fringilla, ligula sed porta fringilla, ligula wisi commodo felis, ut adipiscing felis dui in enim. Suspendisse malesuada ultrices ante. Pellentesque scelerisque augue sit amet urna. Nulla volutpat aliquet tortor. Cras aliquam, tellus at aliquet pellentesque, justo sapien commodo leo, id rhoncus sapien quam at erat. Nulla commodo, wisi eget sollicitudin pretium, orci orci aliquam orci, ut cursus turpis justo et lacus. Nulla vel tortor. Quisque erat elit, viverra sit amet, sagittis eget, porta sit amet, lacus.

### 3.1 Considerações Finais

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam

mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.

# 4 Resultados e Discussão

Praesent facilisis, augue a adipiscing venenatis, libero risus molestie odio, pulvinar consectetur felis erat ac mauris. Nam vestibulum rhoncus quam. Sed velit urna, pharetra eu, eleifend eu, viverra at, wisi. Maecenas ultrices nibh at turpis. Aenean quam. Nulla ipsum. Aliquam posuere luctus erat. Curabitur magna felis, lacinia et, tristique id, ultrices ut, mauris. Suspendisse feugiat. Cras eleifend wisi vitae tortor. Phasellus leo purus, mattis sit amet, auctor in, rutrum in, magna. In hac habitasse platea dictumst. Phasellus imperdiet metus in sem. Vestibulum ac enim non sem ultricies sagittis. Sed vel diam.

## 4.1 ICDAR 2011

O conjunto de imagens de validação do ICDAR 2011 conta com 141 imagens e é importante de constatar que são imagens de relativa baixa resolução para os padrões atuais, onde as maiores contêm cerca de 315 mil pixels. São caracterizadas por estarem no contexto de imagens de anúncios e anexos de emails, com textos primariamente horizontais.

Ao aplicar a solução sobre os exemplos de validação do dataset, em termos de desempenho, todas as imagens foram processadas com sucesso em 36.31 segundos em um ambiente com aceleração gráfica, média de 257 ms por imagem. Para fins comparativos, a mesma tarefa agora executada sem a aceleração gráfica levou 328.44 segundos, cerca de 9 vezes mais tempo, executando na mesma máquina alocada no serviço Paperspace, que consta com 30GB de memória RAM, disponibilidade de processamento gráfico com NVIDIA QUADRO M4000 e processamento CPU baseado em instâncias AWS C4, disponibilizam 8 núcleos, 16 threads do processador Intel Xeon E5-2666, com clock de 2.9 GHz.

Apesar do maior tempo de processamento, o uso da solução em ambientes sem uma placa gráfica compatível ainda não se torna inviável, sobretudo considerando que a média de tempo de processamento por imagem foi de 2.32 segundos. No entanto, fica desencorajado a execução sobre um conjunto muito grande de imagens, já que o processamento corresponde a uma carga muito alta dependendo do hardware onde a solução for executada, já que a execução em CPU alocou aproximadamente 17.4GB de memória, volume maior do que a quantidade total de grande parte dos sistemas mais domésticos.

Para executar o script de avaliação dos resultados da solução, o comando abaixo pode ser executado, fornecendo o caminho para um arquivo .ZIP que contém os arquivos de resultados para cada imagem do conjunto de validação.

```
python script.py -g=gt.zip -s=res_img_.zip
```



Figura 4 – Exemplo de resultado de reconhecimento. Imagem 52 do conjunto de validação.

O resultado da melhor configuração foi o apresentado na Tabela 1. Em suma, a avaliação usa três métricas: Precisão, Recall e F1-Score. A Precisão é a relação dos acertos (verdadeiros positivos) sobre a contagem total de previsões positivas (verdadeiros positivos e falsos positivos), ou seja, dentre as previsões da solução, qual é a porcentagem de acerto. O Recall relaciona a quantidade de acertos com a quantidade total de casos verdadeiros. O F1-Score é basicamente a média harmônica das métricas Precisão e Recall.

```

Calculated!
{
  "precision": 0.7068273092369478,
  "recall": 0.7343532684283728,
  "hmean": 0.7203274215552524,
  "AP": 0
}
  
```

Tabela 1 – Avaliação de resultados sobre a base ICDAR 2011.

Precisão (%)	Recall (%)	F1-Score (%)
70.68	73.44	72.03

Em resumo, isso demonstra que a cada 100 palavras detectadas e processadas, aproximadamente 71 tiveram o texto corretamente extraído. É um resultado bem satisfatório para o trabalho desenvolvido e o nível de simplicidade que foi adotado. Estes números colocariam essa solução em nono lugar entre 18 outros trabalhos submetidos na plataforma de desafios Robust Reading Competition.



Figura 5 – Comparação entre entrada e saída sobre a imagem 5 do set de validação do ICDAR 2011

A Fig. 4 exemplifica as etapas da solução apresentada. A imagem de entrada é processada pela rede de detecção que extrai as regiões de texto regredindo as coordenadas das bounding-boxes. Com isso, essas regiões são então cortadas da imagem original e alimentam a rede de reconhecimento para extração do texto decodificado.

Apesar do resultado satisfatório, o mais interessante é aprofundar não nos acertos, onde o modelo reconheceu as palavras certas, mas sim onde ele errou e identificar limitações e, eventualmente, futuras otimizações.

Um dos casos mais evidentes onde a solução apresentou problemas foi na presença de caracteres especiais. Uma das limitações do modelo de reconhecimento é a lista de caracteres passíveis de reconhecimento, que contempla apenas os caracteres do alfabeto da língua inglesa, de A até Z, adicionado dos dígitos decimais, de 0 até 9. Este dicionário limitado restringe muito a capacidade de detecção em que não são tão difíceis de observar. Caracteres de pontuação, acentos, marcações como o cifrão (\$), entre outros casos, acabam dificultando o acerto do reconhecimento. A Fig. 5 exemplifica esta constatação. Nela, é possível observar que o trecho que representa o preço do item anunciado na imagem 5 do conjunto de validação, que contém cifrão, vírgula e asterisco, não foi reconhecida com sucesso. Apesar de o método aproximar bem os caracteres desconhecidos, reconhecendo o cifrão (\$) como um cinco (5) e o asterisco (\*) como a letra X, o resultado não é satisfatório nesse caso.

Outra dificuldade ficou aparente em imagens de baixa resolução, que acabam contendo regiões de texto ainda menores. Em geral, a regressão das bounding boxes fica um pouco mais grosseira, mas ainda satisfatórias. Entretanto, a maior limitação foi o reconhecimento. As regiões de texto recortadas da imagem original acabam ficando bem pequenas e pixeladas, o que trouxe dificuldades para o modelo pré-treinado CRNN utilizado. A Fig. 6 exemplifica essa dificuldade. Pode-se observar que nenhum dos "títulos" de cada uma das fotos que estão presentes na Fig. 6 foram localizadas com sucesso, mas não obtiveram a igualdade entre predição e gabarito (ficaram marcadas com uma área azul).

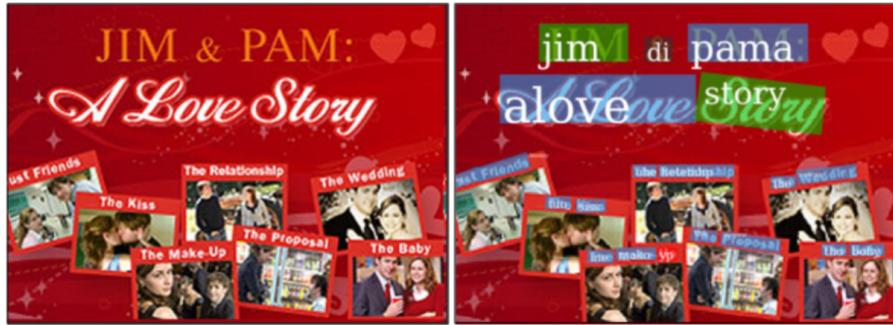


Figura 6 – Exemplo de imagem com instâncias de texto bem pequenas em resolução.  
Imagen 28 do ICDAR 2011

Um outro detalhe que demonstra uma dificuldade mais global do problema de reconhecimento de texto em cenas é a fonte utilizada na frase “A Love Story”, na região superior da Fig. 6. Apesar do reconhecimento ter relativo sucesso nesse exemplo, a localização não conseguiu segmentar corretamente todas as palavras da frase. A generalização da detecção e do reconhecimento para os mais diversos tipos de fontes é um dos principais problemas que motivam o uso de redes neurais profundas para reconhecimento de texto. [6]

O ICDAR 2011 apresenta casos mais complexos de detecção de reconhecimento de texto se comparado ao contexto de aplicações OCR convencionais, que lidam com reconhecimento de texto estruturado, sem grandes variações de fontes e planos de fundo, o que é bastante válido para avaliar a solução. No entanto, ainda não são casos reais de texto em cena, problema que motivou este trabalho. O dataset ICDAR 2013 contém mais exemplos de texto de cena.

#### 4.1.1 ICDAR 2013

O conjunto de imagens de validação do ICDAR 2013 conta com 233 imagens de tamanhos variados. As menores têm cerca de 640 pixels de altura e 480 pixels de comprimento, enquanto as maiores têm dimensões comparáveis à resolução 4K, com 3888 pixels de altura e 2592 pixels de comprimento. Vale ressaltar que as imagens passam por uma redução em resolução ao entrarem no modelo CRAFT, que maximiza a maior dimensão da imagem para 1280 pixels e ajusta a segunda dimensão mantendo a relação de aspecto original, justamente para otimizar o desempenho do método.

Ao aplicar a solução sobre os exemplos de validação do dataset, em termos de desempenho, todas as imagens foram processadas com sucesso em 207.74 segundos em um ambiente com aceleração gráfica, exatamente os mesmos recursos computacionais disponíveis na validação do ICDAR 2011, atingindo uma média de 892 ms por imagem. A média de tempo de processamento por imagem é cerca de 3.5 vezes maior quando comparado ao dataset anterior, ICDAR 2011. As predições não foram executadas sem



Figura 7 – Demonstração de um falso positivo durante a avaliação da solução contra o dataset ICDAR 2013

aceleração gráfica por economia de recursos.

Com base nas mesmas métricas apresentadas na Seção 4.1, durante a discussão sobre os resultados contra o ICDAR 2011, a avaliação da solução no ICDAR 2013 demonstra resultados similares aos vistos na Seção 4.1, disponíveis na Tabela 2. Cerca de 7 acertos a cada 10 previsões. Tendo em vista que é uma base um pouco mais desafiadora, é um resultado novamente satisfatório. Comparando precisão e recall, nota-se um desvio maior, de aproximadamente 6 pontos percentuais, o que indica um maior número de falsos positivos, ou seja, regiões detectadas como regiões de texto que não estão nas anotações de gabarito. Em alguns casos, uma mesma palavra acabou sendo segmentada de maneira errada, em outros, regiões visivelmente de não-texto foram detectadas. A Fig.7 demonstra esse fenômeno.

```
Calculated!
{
    "precision": 0.7043390514631686,
    "recall": 0.7611777535441657,
    "hmean": 0.7316561844863733,
    "AP": 0
}
```

Tabela 2 – Avaliação de resultados sobre a base ICDAR 2013.

Precisão (%)	Recall (%)	F1-Score (%)
70.43	76.11	73.17

Entretanto, as principais dificuldades de detecção e reconhecimento são compartilhadas com o que foi observado durante a avaliação na base ICDAR 2011. Caracteres especiais, pontuação e acentos são limitações conhecidas do reconhecimento.

Por introduzir mais exemplos de imagens verdadeiramente com textos em ambientes naturais, algumas novas dificuldades apareceram. O desafio que provavelmente é mais evidente está relacionado aos eventuais artefatos nas imagens decorrente de reflexos em



Figura 8 – Exemplo de imagem com artefatos que trouxeram dificuldades para a localização correta do texto.



Figura 9 – Exemplo de texto sob vidro, com plano de fundo desafiadores para o reconhecimento.

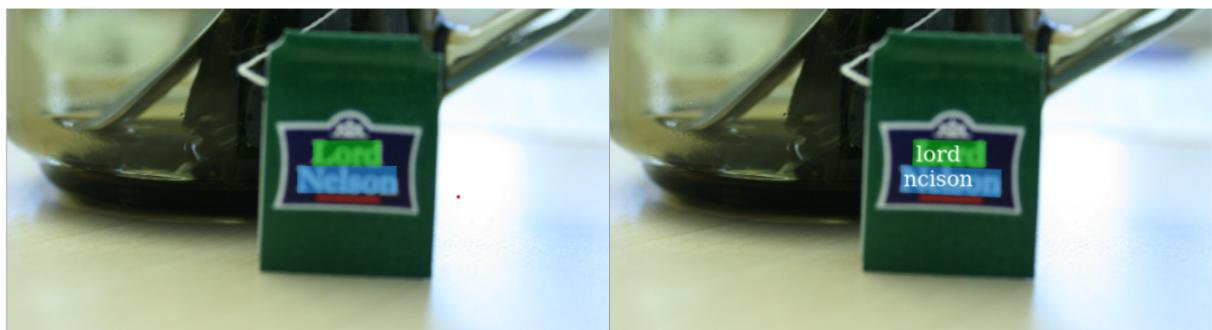


Figura 10 – Exemplo de imagem com texto em desfoco

algumas imagens de texto sob vidro ou com incidência de iluminação de alta intensidade (*flashes* de câmeras fotográficas), onde, em alguns exemplos, a detecção foi sub-ótima (Fig. 8) e em outros, o reconhecimento não foi o melhor (Fig. 9).

Outra dificuldade conhecida de problemas de reconhecimento de texto em cenas está relacionada a palavras em destaque, o que pode ser relativamente comum, dado o contexto de uma imagem com diversos objetos na cena. Apesar de ser um exemplo mais simples, a Fig. 10 demonstra esse caso. A etiqueta na infusão está visivelmente em desfoco e complicou o trabalho do reconhecimento por deixar os caracteres mais ambíguos.

#### 4.1.2 Imagens autorais

Com caráter mais qualitativo, é interessante inferir a qualidade das soluções em imagens fora do contexto de datasets conhecidamente utilizados para treino e validação, para experienciar a qualidade do reconhecimento em imagens do dia-a-dia. Com esse



Figura 11 – Exemplo de imagem autoral onde a solução apresentou dificuldades com texto curvo e estilizado. Textos reconhecidos: “sney” e “intmalakingon”.



Figura 12 – Exemplo de imagem autoral onde uma região de texto muito longa não foi extraída muito bem, principalmente no início da palavra, o que dificultou o reconhecimento. Texto reconhecido: “permmusem”.

objetivo, algumas imagens no álbum pessoal do autor deste trabalho foram selecionadas para passarem pela solução apresentada. Alguns exemplos estão disponíveis abaixo.

Algumas imagens utilizadas foram bastante desafiadoras para a solução, tanto em termos de localização quanto reconhecimento. Imagens com texto bem localizado e visível, alguns letreiros e placas tiveram bons resultados, mas alguns casos de palavras mais longas e curvadas, mesmo que levemente, não tiveram muito sucesso, especialmente se apresentarem caligrafia muito estilizada, como é possível observar na Fig. 13. A respeito de textos curvados, melhorias na solução poderiam melhorar a situação, como a aplicação de soluções para transformar as imagens antes de passar pelo reconhecimento como, por exemplo, uma rede STN (*Spatial Transformer Network* [12]), a fim de obter uma imagem com o mínimo de curvatura possível, facilitando o trabalho de reconhecimento.

Um exemplo foi bastante interessante, pois contrariou as expectativas quanto ao reconhecimento, o que demonstra o potencial que a solução pode apresentar com algumas iterações de melhorias. A Fig. 14 mostra um caso de sucesso na localização e reconhecimento onde o contexto no qual o texto está inserido tem bastante informação e a fonte do texto, principalmente do trecho escrito “Disney’s”, é bastante estilizado, e mesmo assim, o reconhecimento foi bem preciso.

Agora, quando o texto se encontra em situações bastante favoráveis, por exemplo, bem iluminado, sem oclusões, em geral disposto horizontalmente e com caligrafia não rebruscada, a solução atinge o seu potencial máximo, conseguindo localizar e interpretar com muita precisão. A Fig. 15 demonstra esse caso, onde temos um cartaz de boas-vindas do museu do Apartheid, na África do Sul.



Figura 13 – Exemplo de imagem autoral com fontes bastante estilizadas que trouxeram dificuldade tanto para a detecção, quanto para o reconhecimento. Textos reconhecidos: “sey” e “fpan”.



Figura 14 – Exemplo de reconhecimento com sucesso em condições desafiadoras em imagem autoral. Textos reconhecidos: “disneys”, “electrical” e “parade”.



Figura 15 – Exemplo de imagem autoral com alta precisão de reconhecimento. Textos reconhecidos: “nelson”, “rolihlahla” “mandela”, “1918”, “2013”, “hamba”, “kahle”, “madiba”, “ve”, “honour”, “your”, “legacy”, “apartheid”, “museum”.

## 5 Conclusão

Proin non sem. Donec nec erat. Proin libero. Aliquam viverra arcu. Donec vitae purus. Donec felis mi, semper id, scelerisque porta, sollicitudin sed, turpis. Nulla in urna. Integer varius wisi non elit. Etiam nec sem. Mauris consequat, risus nec congue condimentum, ligula ligula suscipit urna, vitae porta odio erat quis sapien. Proin luctus leo id erat. Etiam massa metus, accumsan pellentesque, sagittis sit amet, venenatis nec, mauris. Praesent urna eros, ornare nec, vulputate eget, cursus sed, justo. Phasellus nec lorem. Nullam ligula ligula, mollis sit amet, faucibus vel, eleifend ac, dui. Aliquam erat volutpat.



## Referências

- 1 BAEK, Y. et al. Character region awareness for text detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 9365–9374. Citado 4 vezes nas páginas [5](#), [7](#), [8](#) e [9](#).
- 2 SHI, B.; BAI, X.; YAO, C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 39, p. 2298–2304, 2017. Citado 4 vezes nas páginas [5](#), [9](#), [10](#) e [11](#).
- 3 ZHAO, W.; JIANG, W.; QIU, X. Deep learning for covid-19 detection based on ct images. *Scientific Reports*, v. 11, 2021. Citado na página [1](#).
- 4 Gustav Tauschek. *Reading machine*. US2026330A, 27 maio 1929. Citado na página [1](#).
- 5 RAISI, Z. et al. Text detection and recognition in the wild: A review. *ArXiv*, abs/2006.04305, 2020. Citado na página [2](#).
- 6 LONG, S.; HE, X.; YAO, C. Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision*, v. 129, p. 161–184, 2020. Citado 2 vezes nas páginas [2](#) e [18](#).
- 7 NAVER Corporation. Disponível em: <<https://www.navercorp.com/en/>>. Citado na página [7](#).
- 8 SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015. Citado na página [7](#).
- 9 RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: *MICCAI*. [S.l.: s.n.], 2015. Citado na página [7](#).
- 10 GUPTA, A.; VEDALDI, A.; ZISSERMAN, A. Synthetic data for text localisation in natural images. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2315–2324, 2016. Citado na página [8](#).
- 11 KORNILOV, A. S.; SAFONOV, I. V. An overview of watershed algorithm implementations in open source libraries. *Journal of Imaging*, v. 4, n. 10, 2018. ISSN 2313-433X. Disponível em: <<https://www.mdpi.com/2313-433X/4/10/123>>. Citado na página [8](#).
- 12 JADERBERG, M. et al. Spatial transformer networks. *CoRR*, abs/1506.02025, 2015. Disponível em: <<http://arxiv.org/abs/1506.02025>>. Citado na página [21](#).