



FACULDADE DE TECNOLOGIA DO AMAPÁ - META
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

BETO BATISTA DE SOUZA
WILLIAM KELVEM DE SOUSA PEREIRA

SICOMUV: Sistema de Comunicação Multifuncional com Reconhecimento de Texto e
Assistência por Voz para Inclusão Digital

MACAPÁ – AP
2024

BETO BATISTA DE SOUZA
WILLIAM KELVEM DE SOUSA PEREIRA

SICOMUV: Sistema de Comunicação Multifuncional com Reconhecimento de Texto e
Assistência por Voz para Inclusão Digital

Trabalho de Conclusão de Curso na
Modalidade de Monografia apresentada à
Faculdade de Tecnologia do Amapá - META,
como requisito final à obtenção do grau de
Bacharel em Engenharia da Computação, sob a
orientação do Prof. Ms. Michael Santos de
Oliveira.

MACAPÁ – AP
2024

BETO BATISTA DE SOUZA
WILLIAM KELVEM DE SOUSA PEREIRA

**SICOMUV: Sistema de Comunicação Multifuncional com Reconhecimento de Texto e
Assistência por Voz para Inclusão Digital**

Trabalho de Conclusão de Curso
na Modalidade de Monografia
apresentada à Faculdade de
Tecnologia do Amapá - META,
como requisito final à obtenção do
grau de Bacharel em Engenharia
de Computação.

DATA DE APROVAÇÃO: ____/____/____.

BANCA EXAMINADORA

Prof. Ms. Michael Santos de Oliveira (Orientador)

Prof. XXXXXXXXXXXXXXXXXXXXXXXX (Avaliador 1)

Profª. XXXXXXXXXXXXXXXXXXXXXXXX (Avaliador 2)

AGRADECIMENTOS

Em primeiro lugar, agradecemos a Deus pela força, sabedoria e discernimento, que nos permitiram enfrentar e superar todos os desafios desta jornada acadêmica.

Aos nossos pais e familiares, manifestamos nossa mais profunda gratidão pelo apoio incondicional e confiança depositados em nós, especialmente nos momentos mais difíceis, proporcionando-nos a base sólida necessária para seguir em frente.

Ao nosso orientador, Prof. Me. Michael Santos de Oliveira, expressamos nosso reconhecimento pela paciência, orientação atenciosa e generosidade ao compartilhar seu vasto conhecimento, elementos estes essenciais para a conclusão deste trabalho.

Ao coordenador do curso de Bacharelado em Engenharia da Computação, Prof. Felipe Souza, agradecemos sinceramente pelo suporte constante na resolução de dúvidas e dificuldades que surgiram ao longo do curso, especialmente durante a elaboração deste Trabalho de Conclusão de Curso.

À Profª. Me. Quele Rodrigues, registramos nossos sinceros agradecimentos pela orientação precisa, paciência e pelos valiosos conhecimentos transmitidos ao longo do período acadêmico, bem como pelas intervenções firmes e necessárias que contribuíram significativamente para o nosso crescimento pessoal e profissional.

À nossa colega de turma Maria Gabriele, somos profundamente gratos pela valiosa colaboração durante o processo de estruturação e correção do artigo conforme as normas necessárias. Reconhecemos e agradecemos sua paciência, dedicação, mesmo diante de eventuais divergências, assim como pela troca enriquecedora de conhecimentos ao longo deste percurso.

Aos participantes desta pesquisa, agradecemos imensamente pela disponibilidade e pelas informações fornecidas, fundamentais para a concretização deste estudo.

Por fim, estendemos nossos sinceros agradecimentos a todos que, direta ou indiretamente, contribuíram para que esta importante etapa fosse concluída com sucesso.

SICOMUV: Sistema de Comunicação Multifuncional com Reconhecimento de Texto e Assistência por Voz para Inclusão Digital

Beto Batista de Souza¹
William Kelvem de Sousa Pereira²
Michael Santos de Oliveira³

RESUMO

O projeto SICOMUV propõe o desenvolvimento de uma solução tecnológica para melhorar a acessibilidade de pessoas com deficiência visual, facilitando o acesso a textos impressos e digitais. Este sistema visa enfrentar o desafio do acesso limitado a informações escritas, uma necessidade urgente em um mundo cada vez mais digital. Utilizando tecnologias como o Reconhecimento Óptico de Caracteres (OCR) e redes neurais convolucionais (CNNs), o SICOMUV converte texto em áudio, permitindo a interação em múltiplos idiomas. A metodologia envolve a implementação do Tesseract OCR para extração de texto e a integração de frameworks de aprendizado profundo para aumentar a precisão do reconhecimento. Embora o sistema mostre eficácia em diversos cenários, ele ainda enfrenta limitações em condições de baixa iluminação e com fontes não convencionais, o que impacta sua robustez e adaptabilidade. Os resultados indicam que o SICOMUV atinge seus objetivos de inclusão digital, mas há espaço para melhorias, especialmente na adaptação a ambientes desafiadores. A conclusão destaca que, apesar das limitações, o projeto oferece uma base sólida para futuras inovações, como a incorporação de técnicas avançadas de inteligência artificial para aprimorar o processamento de dados visuais. Assim, o SICOMUV se posiciona como uma solução promissora e inovadora para a comunicação assistiva, reforçando seu compromisso contínuo com a acessibilidade e a inclusão digital.

Palavras-chave: Inclusão digital. Reconhecimento de texto. Assistência por voz. Tradução automática.

¹ Acadêmico do Curso de Bacharelado em Engenharia da Computação da Faculdade de Tecnologia do Amapá – META; E-mail: betobatista354@gmail.com

² Acadêmico do Curso de Bacharelado em Engenharia da Computação da Faculdade de Tecnologia do Amapá – META; E-mail: williamkelvem64@gmail.com

³ Mestre em Computação Aplicada - UFPA, docente do curso de Engenharia de Computação; E-mail: michael.oliveira@tucurui.ufpa.br

**SICOMUV: Multifunctional Communication System with Text Recognition and Voice
Assistance for Digital Inclusion**

Beto Batista de Souza⁴
William Kelvem de Souza Pereira⁵
Michael Santos de Oliveira⁶

ABSTRACT

The SICOMUV project proposes the development of a technological solution to enhance accessibility for visually impaired individuals, facilitating access to printed and digital texts. This system aims to address the challenge of limited access to written information, an urgent need in an increasingly digital world. Utilizing technologies such as Optical Character Recognition (OCR) and Convolutional Neural Networks (CNNs), SICOMUV converts text into audio, enabling interaction in multiple languages. The methodology involves implementing Tesseract OCR for text extraction and integrating deep learning frameworks to improve recognition accuracy. Although the system demonstrates effectiveness in various scenarios, it still faces limitations in low-light conditions and with unconventional fonts, impacting its robustness and adaptability. The results indicate that SICOMUV achieves its goals of digital inclusion, yet there is room for improvement, particularly in adapting to challenging environments. The conclusion highlights that despite these limitations, the project provides a solid foundation for future innovations, such as incorporating advanced artificial intelligence techniques to enhance visual data processing. Thus, SICOMUV positions itself as a promising and innovative solution for assistive communication, reinforcing its ongoing commitment to accessibility and digital inclusion.

Keywords: Digital inclusion. Text recognition. Voice assistance. Automatic translation.

⁴ Academic of the Bachelor's Degree in Computer Engineering at the Faculdade de Tecnologia do Amapá – META; E-mail: betobatista354@gmail.com

⁵ Academic of the Bachelor's Degree in Computer Engineering at the Faculdade de Tecnologia do Amapá – META; E-mail: williamkelvem94@gmail.com

⁶ Master of/in Applied Computing - UFPA, professor of the Computer Engineering course; E- mail: michael.oliveira@tucurui.ufpa.br

LISTA DE ILUSTRAÇÕES

FIGURAS

Figura 1 - Diagrama da Convolução entre uma Matriz e um Núcleo.....	16
Figura 2 - A Arquitetura da CNN Proposta.....	20
Figura 3 - Arquitetura do Modelo TrOCR.....	22
Figura 4 - Conjunto de Dados para Treinamento do Sistema Inteligente de Leitura Automática de Texto.....	24
Figura 5 - IDE do Visual Studio Code.....	25
Figura 6 - Listagem de Modelos e Parametrização para Tradução.....	26
Figura 7 - Bibliotecas Utilizadas no Desenvolvimento do Sistema Inteligente de Leitura Automática de Texto.....	28
Figura 8 - Processamento da Captura, Código para Tradução e Reprodução do Texto em Fala..	29

GRÁFICOS

Gráfico 1 - Distribuição da população de 2 anos ou mais de idade - Brasil.....	11
Gráfico 2 - Nível de Ocupação por Nível de Instrução de Pessoas de 25 Anos ou Mais - Brasil	12

LISTA DE ABREVIATURAS E SIGLAS

CNN	Convolutional Neural Networks (Redes Neurais Convolucionais)
HMM	Hidden Markov Models (Modelos Ocultos de Markov)
HVP-RF	Hierarchical Visual Patterns with Random Forest (Padrões Visuais Hierárquicos com Floresta Aleatória)
IA	Inteligência Artificial
IDE	Integrated Development Environment (Ambiente de Desenvolvimento Integrado)
NIST	National Institute of Standards and Technology (Instituto Nacional de Padrões e Tecnologia)
OCR	Optical Character Recognition (Reconhecimento Óptico de Caracteres)
PLN	Processamento de Linguagem Natural
PNAD	Pesquisa Nacional por Amostra de Domicílios
ReLU	Rectified Linear Unit (Unidade Linear Retificada)
RNN	Recurrent Neural Networks (Redes Neurais Recorrentes)
SICOMUV	Sistema de Comunicação Multifuncional por Voz
SP-RF	Sparse Representation-based Classification with Random Forest (Classificação baseada em Representação Esparsa com Floresta Aleatória)
TrOCR	Transformer-based Optical Character Recognition

SUMÁRIO

1 INTRODUÇÃO.....	10
1.1 MOTIVAÇÃO E JUSTIFICATIVA.....	11
1.2 OBJETIVOS GERAL E ESPECÍFICOS.....	13
1.3 OBJETIVOS ESPECÍFICOS.....	13
2 REFERENCIAL TEÓRICO.....	14
2.1 INTELIGÊNCIA ARTIFICIAL (IA).....	14
2.2 PROCESSAMENTO DE IMAGENS DIGITAIS.....	15
2.3 PROCESSAMENTO DE LINGUAGEM NATURAL.....	17
2.4 RECONHECIMENTO ÓPTICO DE CARACTERES (OCR).....	17
2.5 TÉCNICAS DE SÍNTESE DE VOZ.....	18
2.6 REDES NEURAIS CONVOLUCIONAIS.....	18
2.6.1 Fundamentos e Técnicas Essenciais.....	18
2.6.1.1 RELU (Unidade Linear Retificada).....	18
2.6.1.2 Técnica do Abandono.....	19
2.6.1.3 SP-RF (Classificação baseada em Representação Esparsa com Floresta Aleatória).....	19
2.6.1.4 HVP-RF (Padrões Visuais Hierárquicos com Floresta Aleatória).....	19
2.6.2 Arquitetura e Funcionamento das CNNs.....	19
2.7 TRABALHOS RELACIONADOS.....	20
2.7.1 Artigo 1: Redes Neurais Convolucionais para Reconhecimento de Imagem e Classificação de Documentos.....	20
2.7.2 Artigo 2: TrOCR: Reconhecimento Óptico de Caracteres Baseado em Transformador com Modelos Pré-Treinados.....	21
2.7.3 Análise dos Trabalhos Relacionados.....	22
3 METODOLOGIA.....	24
3.1 FERRAMENTAS UTILIZADAS.....	24
3.2 CONJUNTO DE DADOS.....	25
3.3 BIBLIOTECAS UTILIZADAS.....	26
3.4 IMPLEMENTAÇÃO DO SISTEMA.....	28
3.5 TRADUÇÃO E REPRODUÇÃO EM ÁUDIO.....	29
3.6 AVALIAÇÃO DO DESEMPENHO DO MODELO TREINADO.....	30
3.7 CONSIDERAÇÕES SOBRE O AMBIENTE DE TESTE.....	30
4 ANÁLISE E DISCUSSÃO DOS RESULTADOS.....	32
5 CONSIDERAÇÕES FINAIS.....	33
REFERÊNCIAS BIBLIOGRÁFICAS.....	34

1 INTRODUÇÃO

O Reconhecimento Óptico de Caracteres (OCR) tem emergido como uma tecnologia essencial na era digital, permitindo a extração de informações de textos presentes em documentos físicos e digitais. Com os avanços em técnicas de processamento de imagens e aprendizado de máquina, o OCR tem se tornado cada vez mais preciso e acessível, impulsionando o desenvolvimento de sistemas que facilitam a interação e a comunicação em diversas áreas. A capacidade de converter texto impresso em dados editáveis não apenas melhora a eficiência em ambientes de trabalho, mas também promove a inclusão social, especialmente para pessoas com deficiência visual (SAMPAIO *et al.*, 2023).

A escolha deste tema é motivada pela crescente necessidade de soluções que tornem a informação acessível a um público mais amplo. Dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2022 indicam que aproximadamente 18,6 milhões de pessoas com deficiência vivem no Brasil, evidenciando a urgência de tecnologias que promovam a inclusão e a igualdade de oportunidades (IBGE, 2022). O uso de tecnologias assistivas, como o OCR, pode melhorar significativamente a autonomia e a qualidade de vida dessas pessoas, permitindo-lhes acessar informações e participar ativamente da sociedade (FREITAS *et al.*, 2022).

O sistema proposto neste trabalho, denominado SICOMUV, visa integrar reconhecimento de texto e assistência por voz, contribuindo para a inclusão digital de pessoas com deficiência visual. Ao converter documentos físicos em formato digital, o sistema facilita o reconhecimento de texto para a tradução, além de permitir uma interação mais fluida com a tecnologia. A literatura aponta que a aplicação de redes neurais convolucionais (CNNs) e modelos de aprendizado profundo tem mostrado resultados promissores na melhoria da precisão do reconhecimento de texto, superando limitações de abordagens tradicionais (KANG, *et al.*, 2014; LI *et al.*, 2023). O modelo TrOCR, por exemplo, utiliza uma arquitetura baseada em transformadores que se destaca na tarefa de reconhecimento de texto, alcançando resultados de ponta em diversos benchmarks (WEI *et al.*, 2022).

Apesar dos avanços, desafios como a precisão e a velocidade do reconhecimento, bem como a adaptação a diferentes tipos de fontes e condições de iluminação, ainda precisam ser superados. O SICOMUV busca abordar essas questões, utilizando técnicas de aprendizado de máquina e modelos pré-treinados para melhorar a acurácia e a robustez do reconhecimento de texto (TAPPERT, 2007; SMITH, 2007).

Assim, este trabalho não apenas se alinha às tendências atuais no campo do OCR, mas também busca inovar, propondo soluções tecnológicas que possam contribuir

significativamente para a inclusão digital e a promoção da igualdade de oportunidades para pessoas com deficiência. A implementação de um sistema eficaz de OCR pode, portanto, transformar a forma como as informações são acessadas e utilizadas, promovendo uma sociedade mais inclusiva e informada.

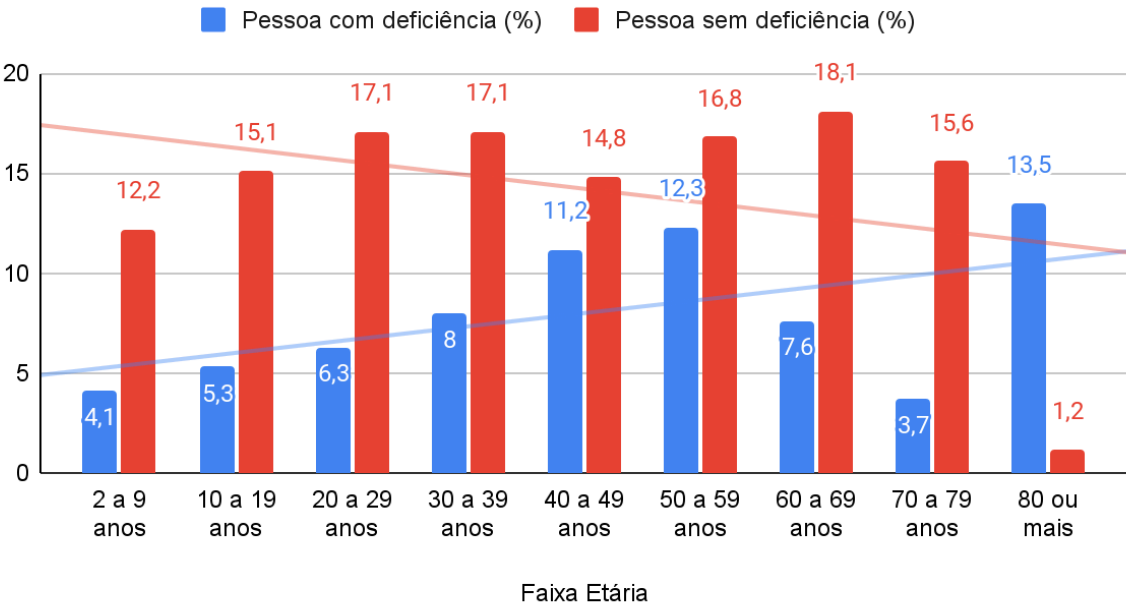
1.1 MOTIVAÇÃO E JUSTIFICATIVA

A inclusão e a promoção dos direitos das pessoas com deficiência são temas de extrema relevância na sociedade contemporânea. Dados recentes da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2022 revelam que aproximadamente 18,6 milhões de pessoas com deficiência vivem no Brasil, representando 8,9% da população com 2 anos ou mais (IBGE, 2022). Essa estatística destaca a necessidade urgente de políticas públicas que abordem as desigualdades enfrentadas por esse grupo.

A pesquisa indica que as mulheres compõem mais da metade da população com deficiência, totalizando 10,7 milhões, o que reforça a importância de considerar a interseccionalidade de gênero nas discussões sobre inclusão. Além disso, o Nordeste é a região com a maior proporção de pessoas com deficiência, atingindo 10,3% da população local, o que sugere a necessidade de estratégias regionais específicas (IBGE, 2022).

Gráfico 1 - Distribuição da população de 2 anos ou mais de idade - Brasil

Pessoas com Deficiência (%) e Pessoas sem Deficiência (%)



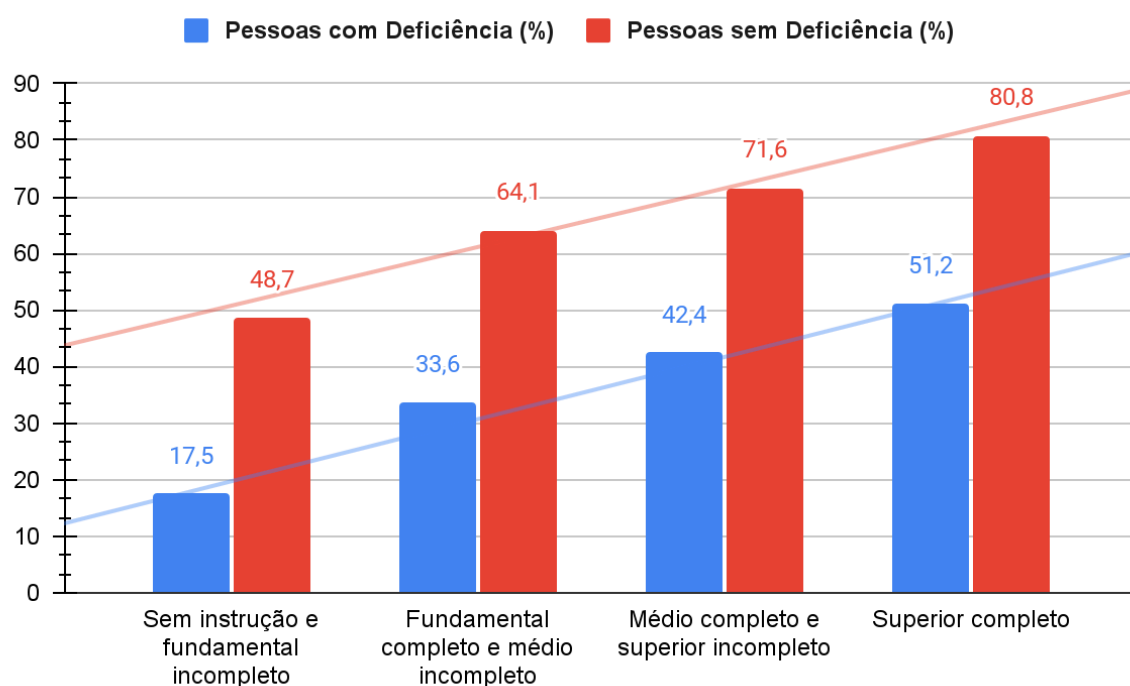
Fonte: Elaborado pelo autor com base nos dados do IBGE (2022).

Outro ponto crucial é a taxa alarmante de analfabetismo entre pessoas com deficiência, que é de 19,5%, em contraste com apenas 4,1% entre pessoas sem deficiência. Essa

disparidade no acesso à educação é um fator motivador para a pesquisa, pois a educação é fundamental para a inclusão social e econômica (IBGE, 2022).

Adicionalmente, apenas 26,6% das pessoas com deficiência estão inseridas no mercado de trabalho, comparado a 60,7% da população sem deficiência. Essa diferença significativa na taxa de emprego, juntamente com a renda média de R\$1.860,00 para pessoas com deficiência e R\$2.690,00 para aqueles sem deficiência, evidencia a urgência de ações que promovam a igualdade de oportunidades (IBGE, 2022).

Gráfico 2 - Nível de Ocupação por Nível de Instrução de Pessoas de 25 Anos ou Mais - Brasil



Fonte: Elaborado pelo autor com base nos dados do IBGE (2022).

A baixa taxa de inclusão de pessoas com deficiência no mercado de trabalho e a alta taxa de analfabetismo evidenciam a necessidade de soluções tecnológicas que promovam a acessibilidade. O sistema proposto neste trabalho visa facilitar a comunicação e a interação de pessoas com deficiência visual, contribuindo para a inclusão social e a igualdade de oportunidades.

Esses dados não apenas ilustram as dificuldades enfrentadas pelas pessoas com deficiência, mas também servem como um chamado à ação para a sociedade e os formuladores de políticas. A compreensão dessas realidades é fundamental para a construção de um futuro mais inclusivo e justo, motivando a realização deste trabalho de pesquisa aplicada.

1.2 OBJETIVOS GERAL E ESPECÍFICOS

Esta monografia tem como objetivo geral desenvolver um sistema inteligente que utiliza técnicas de visão computacional, reconhecimento de caracteres e processamento de linguagem natural para promover acessibilidade às pessoas com deficiência visual, proporcionando uma experiência de leitura inclusiva.

1.3 OBJETIVOS ESPECÍFICOS

- Utilizar uma ferramenta para conversão de texto em áudio.
- Desenvolver e integrar modelos de reconhecimento de caracteres em imagens.
- Implementar um módulo de tradução automática para português e inglês.

2 REFERENCIAL TEÓRICO

Pensar em acessibilidade é reconhecer sua importância como um princípio fundamental para garantir a igualdade e a participação de todas as pessoas na sociedade. No contexto da tecnologia, isso implica na eliminação de barreiras que impedem o acesso e a utilização de informações e serviços por pessoas com diferentes habilidades e características. Da mesma forma, a inclusão digital visa abrir caminhos para promover a igualdade, assegurando que todos possam usufruir dos benefícios proporcionados pelas tecnologias na era digital. Neste sentido, este estudo fundamenta-se em conceitos e teorias relevantes para o desenvolvimento de um sistema inteligente voltado ao auxílio da leitura e à promoção da acessibilidade (SOARES e ALVES, 2023).

2.1 INTELIGÊNCIA ARTIFICIAL (IA)

Segundo Coppin (2010, p. 4), a “Inteligência Artificial é o estudo dos sistemas que agem de um modo que a um observador qualquer pareceria ser inteligente” e complementa que ela “envolve utilizar métodos baseados no comportamento inteligente de humanos e outros animais para solucionar problemas complexos”. Essa definição é boa, porém não ela não engloba toda a Inteligência Artificial (IA), em diversos casos essas técnicas são utilizadas para solução de problemas razoavelmente simples ou complexos (COPPIN, 2010).

Dentro da área de Inteligência Artificial (IA) existem dois tipos, IA forte e IA fraca, abaixo discutidas.

A **IA forte** dispõe de um computador com capacidade suficiente de processamento e fornece a ele inteligência, permitindo assim criar um computador que consiga entender, raciocinar e ser consciente semelhante a um humano, porém muitos filósofos e pesquisadores de IA consideram essa visão improcedente e até insensata de forma que a possibilidade de criar um robô com emoções e consciência genuína é correntemente explorada no campo da ficção científica e é dificilmente considerada como um objetivo da IA (COPPIN, 2010).

Por outro lado, a **IA fraca** é a visão de que o comportamento inteligente pode ser construído e usufruído por computadores para resolver problemas complexos, esse ponto logo visa defender que apenas o fato de um computador agir de forma inteligente não comprova que de fato seja verdadeiramente inteligente no sentido humano (COPPIN, 2010).

Foram discutidas diferenças entre as alegações de IA forte e fraca, porém essa diferença não deve ser confundida com a diferença entre **métodos fortes e métodos fracos**.

Os métodos fracos de IA empregam abordagens como lógica, raciocínio automatizado e outras estruturas genéricas que podem ser adaptadas a uma variedade de problemas, porém não necessariamente incorporam um conhecimento autêntico sobre o

domínio específico em questão, que está sendo resolvido (COPPIN, 2010). Em contrapartida, a resolução de problemas através de **métodos fortes** requer que um sistema possua um amplo conhecimento sobre seu ambiente e os desafios que deve enfrentar.

A capacidade de solucionar problemas por meio de métodos avançados está intrinsecamente ligada aos métodos considerados menos avançados, pois um sistema que possua conhecimento, mas não tenha uma metodologia adequada para lidar com esse conhecimento, torna-se ineficaz (COPPIN, 2010).

2.2 PROCESSAMENTO DE IMAGENS DIGITAIS

Segundo Sá (2021, p. 6):

O processamento de imagens digitais é a base da visão computacional, ou seja, a forma que sistemas computacionais interpretam e processam imagens. Por ser algo totalmente natural aos seres humanos, não se percebe como as imagens são ricas em volume de dados e como a manipulação delas podem gerar soluções e simplificar processos em computação e robótica.

Seguindo essa linha de raciocínio, essas imagens processadas tem uma formação específica, uma composição de cores e intensidade que representam uma cena. O olho humano capta os raios de luz que estão no ambiente, sejam eles originados por alguma fonte, refletidos ou filtrados em outros objetos, muitas vezes, trata-se de uma combinação dessas formas (QUEIROZ, 2006 apud SÁ, 2021, p. 6).

É necessário estabelecer os princípios fundamentais da convolução, módulo de kernel, morfologia, e outros conceitos relacionados. Essas competências são empregadas em todas as etapas de manipulação de imagens, desempenhando funções cruciais na manipulação essencial, extração de características e pré-processamento (SÁ, 2021).

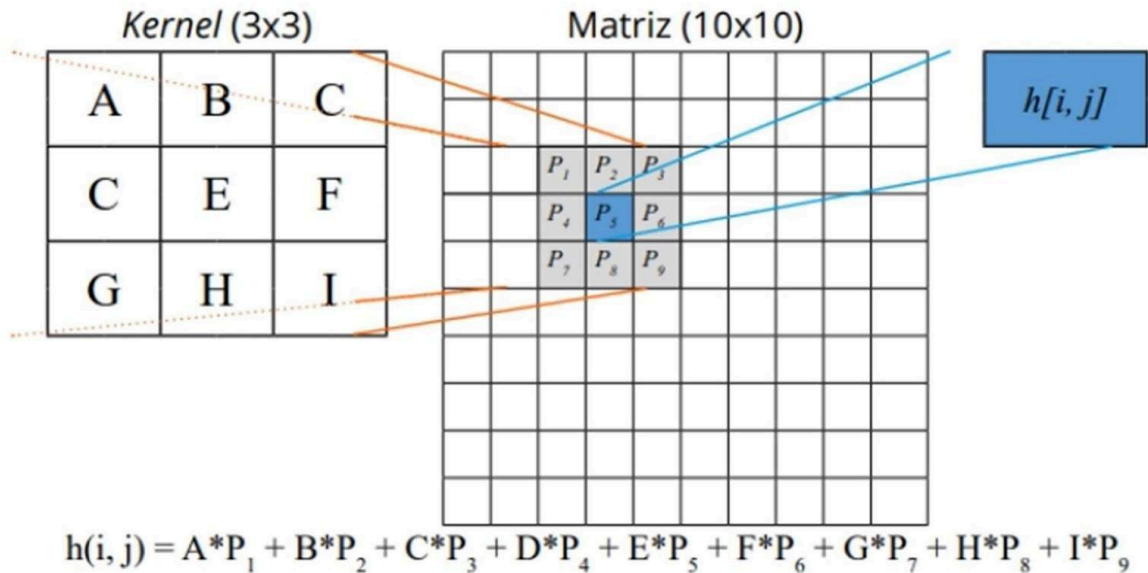
Essas imagens são aplicadas como preparação indispensável antes da utilização de modelos de inteligência artificial, reconhecimento de padrões e visão robótica de forma ampla. As imagens podem ser manipuladas utilizando técnicas algébricas devido à sua natureza numérica, representadas por matrizes de valores (SÁ, 2021).

A convolução é a técnica principal utilizada nesse contexto, envolvendo a aplicação de um kernel sobre a imagem. O kernel pode ser chamado de matriz convolucional, elemento estruturante ou máscara, mas sua descrição mais precisa é a de uma pequena matriz. (SÁ, 2021).

Os elementos do kernel podem ter valores diversos, inclusive fracionários e negativos, mas o resultado do filtro depende da forma como ele é aplicado. Apesar de parecer complexo,

o processo de convolução envolve operações aritméticas simples. O símbolo utilizado para representar a convolução é "X" ou "*" (asterisco) (SÁ, 2021).

Figura 1 - Diagrama da Convolução entre uma Matriz e um Núcleo



FONTE: SÁ (2021, p. 23).

A convolução é uma operação que envolve a soma das multiplicações dos valores das matrizes quando o pixel de destino, como o P5, é alinhado com o centro do kernel, representado por E. Para garantir a validade da convolução, é necessário que o kernel possua um número ímpar de pixels, o que garante a existência de um centro. Não há um tamanho mínimo ou máximo definido para o kernel, mas os tamanhos mais comuns são 3x3 e 5x5 (SÁ, 2021).

Assim, pode-se concluir que a nova matriz resultante da convolução utiliza os valores dos pixels vizinhos. Ela calcula a soma das multiplicações dos pixels correspondentes na imagem principal e cria uma nova imagem com esses valores (SÁ, 2021).

Essa técnica também pode ser aplicada em imagens coloridas, com múltiplos canais, basta aplicar a convolução nas matrizes independentes correspondentes a cada canal. Os softwares gráficos também utilizam a técnica de convolução em seus filtros, permitindo que os usuários insiram seus próprios kernels para criar filtros personalizados (SÁ, 2021).

No entanto, é importante destacar que quando o kernel é aplicado às bordas da imagem original, alguns valores não podem ser calculados corretamente, pois o kernel excede a matriz original. Para mitigar esse problema, são utilizadas técnicas como extensão (copiando o valor do pixel mais próximo), repetição (usando o valor do pixel oposto), espelhamento (usando o valor do pixel oposto em uma das direções), corte (ignorando as

operações com pixels ausentes na matriz principal) e corte do núcleo (ignorando os valores do kernel que excedem os limites) (SÁ, 2021).

Geralmente, os filtros são aplicados várias vezes e, muitas vezes, reaplicados após outros filtros, a fim de obter o efeito desejado na imagem (SÁ, 2021).

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural (PLN) é uma área de pesquisa dentro da inteligência artificial que se dedica a desenvolver métodos computacionais para o tratamento de linguagem humana, buscando capacitar as máquinas a processar e gerar textos de forma semelhante à comunicação humana, abordando tanto a linguagem escrita quanto falada (CASELI, NUNES e PAGANO, 2023). Segundo Jurafsky e Martin (2020), o PLN abrange tarefas como análise de sentimentos, tradução automática, resumo de texto e geração de diálogos, entre outras.

O PLN é um campo multidisciplinar que combina conhecimentos da linguística, ciência da computação e estatística. Conforme destacam Manning e Schütze (1999), especialistas em PLN, as técnicas utilizadas envolvem desde a construção de modelos estatísticos até o uso de algoritmos de aprendizado de máquina. O objetivo é permitir que as máquinas processem e compreendam a linguagem de forma mais eficiente e precisa, a fim de fornecer respostas relevantes e úteis para os usuários.

2.4 RECONHECIMENTO ÓPTICO DE CARACTERES (OCR)

De acordo com Campos e Varma (2009), o Reconhecimento Óptico de Caracteres (OCR) é uma tecnologia que visa extrair informações de texto a partir de imagens ou documentos digitalizados. Ela desempenha um papel fundamental na automação de processos que envolvem a digitalização e a análise de documentos impressos.

Como mencionado por Tappert (2007), um especialista renomado na área, o OCR permite que os computadores convertam caracteres impressos em texto editável, facilitando a busca, a indexação e a manipulação dessas informações.

O OCR utiliza técnicas de processamento de imagem, aprendizado de máquina e reconhecimento de padrões para identificar e interpretar os caracteres presentes em uma imagem. Conforme destaca Smith (2007), outro autor de referência nesse campo, o OCR envolve etapas como segmentação dos caracteres, reconhecimento dos padrões e pós-processamento para melhorar a precisão dos resultados. Com o avanço da tecnologia, o OCR tornou-se uma ferramenta essencial em diversas áreas, como digitalização de documentos, automação de escritórios e até mesmo acessibilidade para pessoas com deficiência visual.

No entanto, o OCR ainda enfrenta desafios, como a correta interpretação de caracteres com ruídos, fontes diferentes ou baixa resolução. Como ressalta Smith (2007), a melhoria contínua dos algoritmos e o treinamento dos modelos com grandes conjuntos de dados têm contribuído para o aprimoramento da precisão do OCR ao longo dos anos. Apesar das dificuldades, o reconhecimento óptico de caracteres desempenha um papel crucial na transformação digital e na otimização de fluxos de trabalho, tornando-se uma tecnologia cada vez mais indispensável em diversos setores.

2.5 TÉCNICAS DE SÍNTESE DE VOZ

As técnicas de síntese de voz são métodos empregados para gerar fala artificial a partir de um texto escrito. Essas técnicas têm avançado significativamente ao longo dos anos, permitindo que as máquinas sejam capazes de reproduzir vozes humanas de forma cada vez mais natural e expressiva (STORY, 2019).

Uma das abordagens amplamente utilizadas na síntese de voz é a síntese concatenativa, que consiste em combinar unidades de fala pré-gravadas para formar palavras e frases. De acordo com Klatt (1987), um pesquisador pioneiro nesse campo, a síntese concatenativa oferece uma alta qualidade de fala, mas requer um grande banco de dados de unidades de fala gravadas. Já a síntese de voz por meio de técnicas de modelagem estatística, como os Modelos Ocultos de Markov (HMMs), tem sido explorada com sucesso.

2.6 REDES NEURAIIS CONVOLUCIONAIS

As Redes Neurais Convolucionais (CNNs) revolucionaram o campo do aprendizado profundo, especialmente no processamento de imagens. Conforme descrito por LeCun *et al.* (2015), as CNNs são arquiteturas especializadas em processar dados com uma topologia estruturada, sendo particularmente eficazes no reconhecimento de padrões visuais diretamente de imagens pixelizadas, minimizando a necessidade de pré-processamento.

2.6.1 Fundamentos e Técnicas Essenciais

2.6.1.1 RELU (Unidade Linear Retificada)

A função de ativação ReLU, introduzida por Nair e Hinton (2010), é definida matematicamente como:

$$f(x) = \max(0, x)$$

Suas principais vantagens incluem a redução do problema de desaparecimento do gradiente, maior eficiência computacional, convergência acelerada durante o treinamento e esparsidade na ativação. Goodfellow *et al.* (2016) destacam que a ReLU supera significativamente as funções de ativação tradicionais, como a sigmoide e a tangente hiperbólica, especialmente em redes profundas.

2.6.1.2 Técnica do Abandono

O Dropout, proposto por Srivastava *et al.* (2014), é uma técnica de regularização que reduz o overfitting ao alternar aleatoriamente neurônios com uma probabilidade p (p = taxa de abandono) durante o treinamento. Essa técnica cria sub-redes diferentes a cada iteração, promovendo a generalização do modelo. Durante a inferência, todos os neurônios são utilizados, e as saídas são multiplicadas por $1-p$ para compensação. Ao forçar a rede a aprender representações mais robustas, o dropout ajuda a evitar que o modelo se ajuste excessivamente aos dados de treinamento.

2.6.1.3 SP-RF (Classificação baseada em Representação Esparsa com Floresta Aleatória)

A técnica SP-RF, desenvolvida por Kumar e Doermann (2013), combina Representação Esparsa e Floresta Aleatória. A Representação Esparsa codifica características de maneira eficiente, reduzindo a dimensionalidade dos dados e capturando padrões essenciais, enquanto a Floresta Aleatória utiliza um conjunto de árvores de decisão, classificando dados de maneira robusta e resistente a ruídos.

2.6.1.4 HVP-RF (Padrões Visuais Hierárquicos com Floresta Aleatória)

A técnica HVP-RF, proposta por Kumar e Doermann (2013), extrai padrões visuais hierárquicos de imagens usando uma análise em múltiplas escalas para capturar relações estruturais. Utiliza-se também Florestas Aleatórias para classificação, o que aumenta a robustez às variações estruturais e melhora a precisão.

2.6.2 Arquitetura e Funcionamento das CNNs

Os componentes principais da Arquitetura e Funcionamento das CNNs são as Camadas Convolucionais que realizam a remoção de características por meio da aplicação de grãos na entrada, permitindo o compartilhamento de configurações e a preservação da informação espacial, as Camadas de Pooling que reduzem a dimensionalidade, fornecem invariância a pequenas transformações e resumindo características importantes e as Camadas Totalmente Conectadas que integram as características extraídas para mapear as classes-alvo e realizar a classificação final.

Segundo He *et al.* (2016), o processo de treinamento de CNNs envolve uma fase de propagação direta, em que as imagens são processadas camada a camada, e o cálculo da perda é realizada a comparação entre a saída da rede com os valores reais esperados, utilizando uma função de perda apropriada. Em seguida, ocorre a retropropagação, onde os gradientes da função de perda em relação aos pesos da rede são calculados. Esses gradientes são então usados para ajustar os pesos através de métodos de otimização, como o gradiente descendente. O objetivo é minimizar a perda e melhorar a precisão do modelo.

2.7 TRABALHOS RELACIONADOS

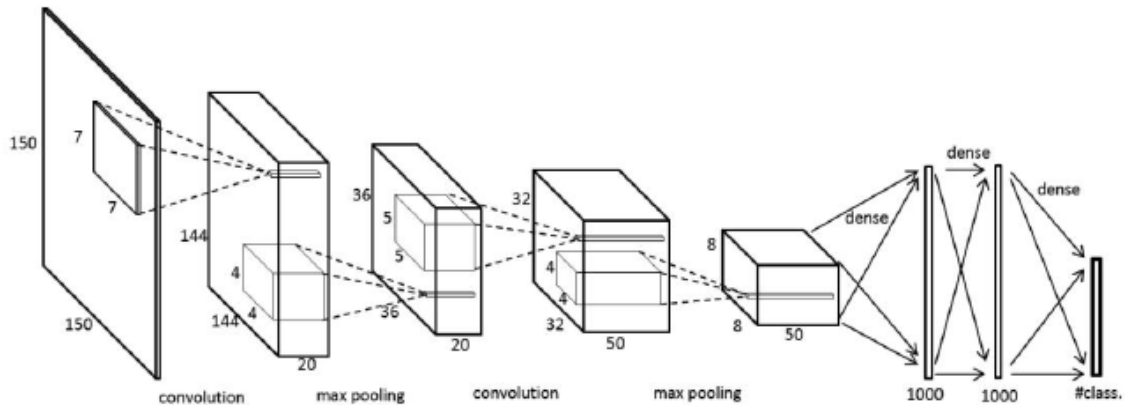
Foram revisados estudos sobre o uso de Inteligência Artificial (IA) para reconhecimento de caracteres e sistemas de conversão de texto em áudio, com foco em técnicas de OCR usando redes neurais convolucionais (CNNs) e recorrentes (RNNs), além de soluções de acessibilidade como descrição de imagens e navegação por voz.

2.7.1 Artigo 1: Redes Neurais Convolucionais para Reconhecimento de Imagem e Classificação de Documentos

Explora uma abordagem inovadora para a classificação de imagens de documentos utilizando CNNs. O foco é superar as limitações das abordagens anteriores, que dependiam de características manuais, propondo o uso de CNNs para aprender automaticamente abstrações de estrutura hierárquica e relações espaciais nas imagens. Isso permite uma classificação mais eficiente de diferentes tipos de documentos (KANG *et al.*, 2014).

A metodologia do estudo envolve o uso de CNNs, que segundo Kang *et al.* (2014) são eficazes em descobrir correlações espaciais locais e reforçar padrões de conectividade entre neurônios em camadas adjacentes. O método utilizado é demonstrado na figura seguinte.

Figura 2 - A Arquitetura da CNN Proposta



Fonte: KANG *et al.*, 2014.

A metodologia do estudo envolve o uso de CNNs, que segundo KANG *et al.* (2014), são eficazes em descobrir correlações espaciais locais e reforçar padrões de conectividade entre neurônios em camadas adjacentes. O processo de classificação é realizado em duas etapas principais:

Pré-processamento: as imagens de documentos são reduzidas e normalizadas para um tamanho de 150×150 pixels, preservando a estrutura geral do layout.

Arquitetura da Rede: a arquitetura da CNN é composta por várias camadas convolucionais e totalmente conectadas, permitindo a aprendizagem hierárquica dos recursos diretamente dos pixels das imagens.

Os experimentos realizados em conjuntos de dados do mundo real, como o conjunto de dados de litígio de tabaco e o conjunto de formulários fiscais NIST, demonstraram que a abordagem proposta alcançou alta precisão na classificação de diferentes gêneros documentais. A CNN superou métodos anteriores, como SP-RF e HVP-RF, com uma precisão média de 65,37% em um dos conjuntos de dados.

O artigo contribui para o campo da classificação de documentos ao propor uma abordagem que aprende recursos diretamente dos pixels, eliminando a necessidade de características manuais, demonstrando a eficácia das CNNs na captura de padrões hierárquicos e na classificação de layouts complexos de documentos. Outra contribuição significativa é a abordagem em conjuntos de dados desafiadores, mostrando que a CNN pode lidar com variações significativas nos layouts de documentos.

A principal inovação do artigo é a utilização de CNNs para aprender automaticamente a estrutura hierárquica dos documentos, permitindo uma classificação mais precisa e eficiente, o uso de ReLU e dropout também melhora o desempenho e a velocidade do treinamento.

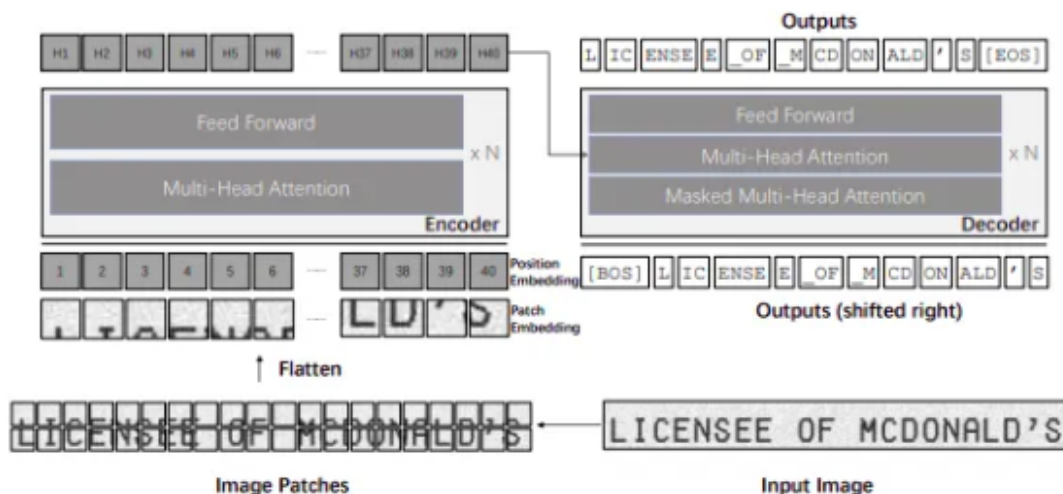
Apesar dos resultados promissores, a abordagem pode ser limitada em cenários onde os documentos apresentam layouts muito variados ou quando se trata de documentos manuscritos, onde as características visuais podem não ser capturadas adequadamente.

O artigo sugere que futuras pesquisas poderiam explorar a aplicação de técnicas de aprendizado de máquina em documentos com layouts mais complexos e diversificados, além de investigar a eficácia da abordagem em documentos manuscritos e não estruturados.

2.7.2 Artigo 2: TrOCR: Reconhecimento Óptico de Caracteres Baseado em Transformador com Modelos Pré-Treinados

Este trabalho introduz uma nova abordagem ao utilizar a arquitetura de Transformer para o reconhecimento de texto, rompendo com o uso tradicional de Redes Neurais Convolucionais (CNNs). Isso reflete uma tendência crescente no aprendizado profundo, que explora os Transformers em tarefas tanto de visão computacional quanto de processamento de linguagem natural (LI *et al.*, 2021). A figura a seguir nos mostra a arquitetura utilizada.

Figura 3 - Arquitetura do Modelo TrOCR



Fonte: WEI *et al* (2022).

Assim como o artigo 1 que visa melhorar a precisão de reconhecimento de padrões visuais, o TrOCR inova ao substituir as CNNs por Transformers. Além disso, o TrOCR obteve resultados superiores em reconhecimento de texto impresso, manuscrito e em cenas, como outros artigos que se concentram em melhorar a precisão em tarefas de visão computacional.

Uma diferença significativa é a eliminação das CNNs, que formam a base dos outros artigos para a extração de características. O TrOCR, ao optar por uma arquitetura completamente baseada em Transformers, traz mais flexibilidade para diferentes cenários de reconhecimento de texto. Também utiliza uma arquitetura de codificador-decodificador, diferente das combinações entre CNNs e RNNs dos outros artigos, simplificando o modelo e suas implementações.

Outro ponto de destaque é o foco em reconhecimento multilíngue, uma possibilidade não abordada explicitamente nos outros artigos, o que confere ao TrOCR uma vantagem para aplicações globais.

O TrOCR se diferencia dos demais ao inovar no uso de Transformers e ao eliminar as CNNs em reconhecimento de texto, alinhando-se com as tendências atuais de pesquisa. Além de contribuir significativamente para o avanço da tecnologia OCR, ele oferece potenciais aplicações no desenvolvimento de sistemas assistivos, como o SICOMUV, promovendo inclusão digital através do reconhecimento de texto e assistência por voz.

2.7.3 Análise dos Trabalhos Relacionados

Os dois artigos analisados abordam o tema do reconhecimento óptico de caracteres (OCR) e suas aplicações em contextos assistivos e de classificação de documentos, mas com

enfoques e metodologias distintas. O artigo "Redes Neurais Convolucionais para Classificação de Imagens de Documentos" propõe uma abordagem que utiliza CNNs para aprender automaticamente a estrutura hierárquica de documentos, enfatizando a eficácia das CNNs em capturar padrões espaciais e hierárquicos. Já o artigo "TrOCR: Reconhecimento Óptico de Caracteres Baseado em Transformador com Modelos Pré-Treinados" apresenta uma abordagem inovadora ao utilizar Transformers, eliminando a dependência de CNNs e RNNs, e alcançando resultados de ponta em reconhecimento de texto impresso, manuscrito e de cena.

As semelhanças entre os artigos incluem o uso de aprendizado de máquina para melhorar a precisão do reconhecimento de texto e o compromisso em atender a necessidades específicas, como a inclusão de pessoas com deficiência. No entanto, as abordagens metodológicas diferem substancialmente. Os dois artigos propõem e validam arquiteturas específicas de redes neurais, sendo que o primeiro se concentra em CNNs para classificação de documentos e o segundo utiliza Transformers para reconhecimento de texto, inovando ao eliminar a dependência das arquiteturas convencionais.

Neste contexto foi possível observar que existem lacunas na literatura que inclui a necessidade de mais pesquisas sobre a eficácia do reconhecimento de textos manuscritos em diferentes estilos de escrita, especialmente em contextos multilíngues. Embora os artigos se destaquem pela predominância de estudos focados em deficiências visuais, há uma falta de atenção a outras deficiências, como auditivas e cognitivas, que poderiam se beneficiar de tecnologias assistivas. Além disso, a exploração de diferentes algoritmos de aprendizado de máquina, além das redes neurais, poderia ser uma área promissora para futuras investigações, visando diversificar as abordagens e melhorar a eficácia das soluções propostas.

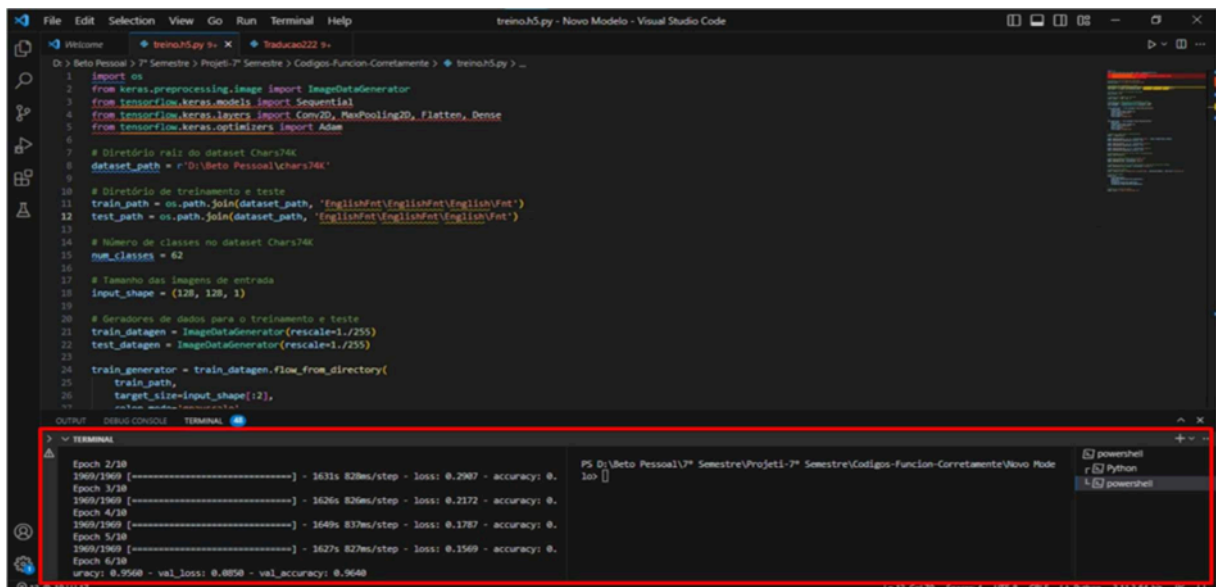
O SICOMUV se propõe a ir além das limitações identificadas nos trabalhos mencionados anteriormente, como a dificuldade do Tesseract em reconhecer letras cursivas e o foco restrito a deficiências visuais. O SICOMUV propõe integrar o reconhecimento de texto manuscrito com assistência por voz utilizando modelos de IA pré-treinados, explorar áreas inexploradas, promovendo uma inclusão digital mais ampla para pessoas com diferentes tipos de deficiência. Além disso, ao utilizar aprendizado de máquina para personalizar o reconhecimento de texto a diferentes estilos de escrita, o projeto aborda diretamente as lacunas mencionadas na literatura, com o objetivo de melhorar a acurácia e a eficácia do sistema.

3 METODOLOGIA

Para o desenvolvimento deste sistema, foi feita uma pesquisa aplicada através da realização de uma revisão bibliográfica detalhada, explorando trabalhos acadêmicos, artigos científicos e livros relevantes na área de reconhecimento de caracteres em imagens.

Foram criados dois códigos principais como parte deste trabalho. O primeiro código é responsável pelo treinamento de um modelo de rede neural convolucional usando a biblioteca Keras e o conjunto de dados Chars74K. Esse modelo treinado será utilizado posteriormente pelo código principal para auxiliar no reconhecimento de caracteres em tempo real. Abaixo, na figura 4 encontra-se o conjunto de dados para treinamento do modelo de rede neural convolucional, sua estruturação e os resultados obtidos em tempo real durante o treinamento.

Figura 4 - Conjunto de Dados para Treinamento do Sistema Inteligente de Leitura Automática de Texto



```
1 import os
2 from keras.preprocessing.image import ImageDataGenerator
3 from tensorflow.keras.models import Sequential
4 from tensorflow.keras.layers import Conv2D, MaxPooling2D, Flatten, Dense
5 from tensorflow.keras.optimizers import Adam
6
7 # Diretório raiz do dataset Chars74K
8 dataset_path = r'D:\Beto Pessoal\chars74k\
9
10 # Diretório de treinamento e teste
11 train_path = os.path.join(dataset_path, 'EnglishFont\EnglishFont\English\Fnt')
12 test_path = os.path.join(dataset_path, 'EnglishFont\EnglishFont\English\Fnt')
13
14 # Número de classes no dataset Chars74K
15 num_classes = 42
16
17 # Tamanho das imagens de entrada
18 input_shape = (128, 128, 1)
19
20 # Geradores de dados para o treinamento e teste
21 train_datagen = ImageDataGenerator(rescale=1./255)
22 test_datagen = ImageDataGenerator(rescale=1./255)
23
24 train_generator = train_datagen.flow_from_directory(
25     train_path,
26     target_size=input_shape[:2],
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
```

Terminal Output:

```
Epoch 2/30
1969/1969 [=====] - 1631s 828ms/step - loss: 0.2907 - accuracy: 0.
Epoch 3/30
1969/1969 [=====] - 1626s 826ms/step - loss: 0.2372 - accuracy: 0.
Epoch 4/30
1969/1969 [=====] - 1649s 837ms/step - loss: 0.1787 - accuracy: 0.
Epoch 5/30
1969/1969 [=====] - 1627s 827ms/step - loss: 0.1569 - accuracy: 0.
Epoch 6/30
1969/1969 [=====] - 1627s 827ms/step - loss: 0.1569 - accuracy: 0.
uracy: 0.9568 - val_loss: 0.8858 - val_accuracy: 0.9648
```

Fonte: Autoria própria (2023).

O segundo código é o núcleo do sistema proposto. Ele utiliza a biblioteca OpenCV para captar imagens da câmera, realiza o pré-processamento dessas imagens, faz a predição dos caracteres utilizando o modelo treinado e, por fim, converte o texto resultante em áudio, utilizando a biblioteca pyttsx3, que é uma síntese de voz offline para Python.

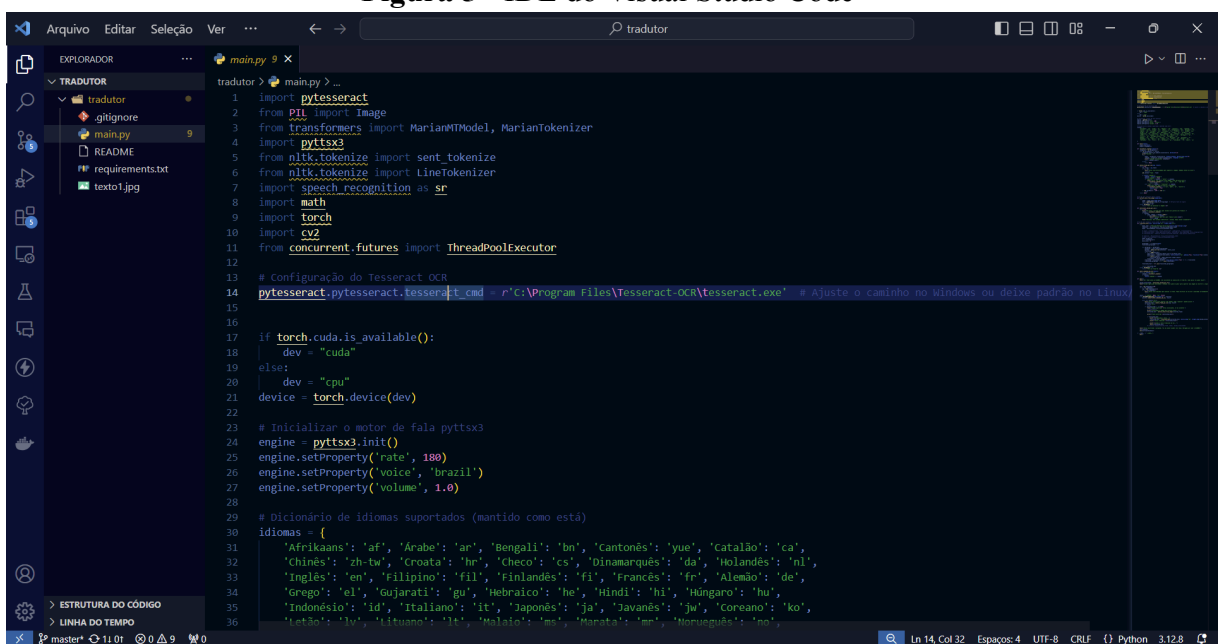
3.1 FERRAMENTAS UTILIZADAS

Para o desenvolvimento do sistema de reconhecimento de caracteres, utilizou-se o ambiente de programação Visual Studio Code. Esse ambiente fornece uma plataforma integrada e de fácil utilização, com recursos avançados de edição de código, depuração e controle de versão, o que facilitou o processo de desenvolvimento do projeto.

A linguagem de programação escolhida para implementar o sistema foi o Python (versão 3.12). Conforme discutido por Al Sweigart no livro "Making Games with Python & Pygame", Python é uma linguagem de programação extremamente versátil, frequentemente utilizada em áreas como aprendizado de máquina e visão computacional. Sua sintaxe clara e concisa, combinada com a ampla disponibilidade de bibliotecas, torna o Python uma escolha ideal para projetos nessas áreas.

Segue abaixo a figura 5 para demonstrar a IDE de desenvolvimento Visual Studio Code que foi mencionada acima.

Figura 5 - IDE do Visual Studio Code



FONTE: Elaborado pelo autor (2024).

3.2 CONJUNTO DE DADOS

Ao realizar o treinamento de um modelo de rede neural convolucional, a escolha do conjunto de dados desempenha um papel crucial na capacidade de generalização do modelo. Nesse contexto, o Chars74K se destaca como um conjunto de dados valioso.

O Chars74K é um conjunto de dados diversificado, composto por caracteres provenientes de uma variedade de fontes e estilos. Essa diversidade é essencial para o treinamento de modelos de reconhecimento óptico de caracteres (OCR), especialmente quando combinados com ferramentas como o Tesseract OCR, configurado como mostrado no trecho de código.

No exemplo de configuração do Tesseract OCR em Python, o caminho para o executável do Tesseract foi ajustado para sistemas Windows:

```
pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-OCR\tesseract.exe'
```

Esse ajuste permite que o modelo utilize o Tesseract para extrair informações visuais diretamente das imagens, proporcionando uma base sólida para reconhecimento de caracteres.

Conforme mostrado na figura 4, ao combinar o Chars74K, que contém caracteres em diversas fontes e estilos, com ferramentas poderosas como o Tesseract OCR, o modelo pode capturar características essenciais que são comuns a diferentes tipos de caracteres. Como resultado, o modelo consegue generalizar e reconhecer caracteres mesmo em contextos ou fontes não vistos durante o treinamento.

Portanto, a configuração do Tesseract OCR junto com o Chars74K potencializa a capacidade de reconhecimento de caracteres em cenários variados, destacando a importância de um bom conjunto de dados e uma configuração apropriada para o sucesso do treinamento.

Como demonstrado na figura 6 logo abaixo, essa capacidade de aprendizado abrangente permite que o modelo reconheça e interprete corretamente os caracteres, mesmo quando são apresentados em fontes ou estilos que não foram explicitamente incluídos no conjunto de treinamento.

Figura 6 - Listagem de Modelos e Parametrização para Tradução



```
161
162 def translate_text(text, source_lang="en", target_lang="pt"):
163     try:
164         model_name = helsinki_models.get((source_lang, target_lang))
165         if not model_name:
166             return "Desculpe, não tenho um modelo disponível para essa tradução."
167
168         tokenizer = MarianTokenizer.from_pretrained(model_name)
169         model = MarianMTModel.from_pretrained(model_name)
170         model.to(device)
171         lt = LineTokenizer()
172         batch_size = 8
173
174         paragraphs = lt.tokenize(text)
175         translated_paragraphs = []
176
177         for paragraph in paragraphs:
178             sentences = sent_tokenize(paragraph)
179             batches = math.ceil(len(sentences) / batch_size)
180             translated = []
181             for i in range(batches):
182                 sent_batch = sentences[i*batch_size:(i+1)*batch_size]
183                 model_inputs = tokenizer(sent_batch, return_tensors="pt", padding=True, truncation=True).to(device)
184                 with torch.no_grad():
185                     translated_batch = model.generate(model_inputs)
186                 translated = [tokenizer.decode(t, skip_special_tokens=True) for t in translated_batch]
187                 translated_paragraphs.append(" ".join(translated))
188
189             translated_text = "\n".join(translated_paragraphs)
190
191             return translated_text
192     except Exception as e:
193         return f"Erro na tradução: {e}"
194
195 def obter_comandos_de_voz(resultado):
196     while not resultado["file"]:
197         comando = reconhecer_comando()
198         if comando:
199             resultado["comando"] = comando
200
201 def main():
202     nltk.download('punkt') # Certifique-se de que o recurso 'punkt' está disponível
203     falar("Olá! Bem-vindo ao SICOMV, seu assistente de comunicação e tradução.")
204
205     idioma_selecionado = selecionar_idioma_por_voz()
```

FONTE: Autoria própria (2024).

3.3 BIBLIOTECAS UTILIZADAS

No desenvolvimento do sistema, diversas bibliotecas foram utilizadas, cada uma desempenhando um papel crucial em diferentes aspectos do projeto.

As bibliotecas que foram utilizadas para aprendizado de máquina e processamento de dados foram o TensorFlow que é uma biblioteca de aprendizado de máquina e redes neurais utilizada para construir e treinar o modelo de reconhecimento de caracteres, o Numpy que é

essencial para a manipulação de arrays e operações matemáticas, processando os dados de entrada do modelo e o Torch que é outra fundamental para aprendizado profundo, oferecendo suporte para a construção de modelos complexos.

Para visão computacional e processamento de imagens foram utilizadas o OpenCV, que é uma biblioteca de visão computacional utilizada para capturar frames da câmera e realizar o pré-processamento das imagens e o Pillow, que adiciona capacidades de processamento de imagem ao Python, permitindo manipulação avançada de imagens.

Foram utilizadas para o processamento de linguagem natural o NLTK, que é uma biblioteca para processamento de linguagem natural, enquanto transformers é usada para implementar modelos de aprendizado profundo baseados em transformadores e o Tokenizers, sacremoses, e sentencepiece, foram utilizadas para tokenização de texto, essenciais para preparar dados textuais para análise.

Para a geração áudio e síntese de fala foi utilizado o Pyttsx3 que gera arquivos de áudio a partir do texto traduzido e para reprodução de áudio e o PyAudio, que facilita a reprodução e gravação de áudio, enquanto SpeechRecognition oferece funcionalidades para reconhecimento de voz.

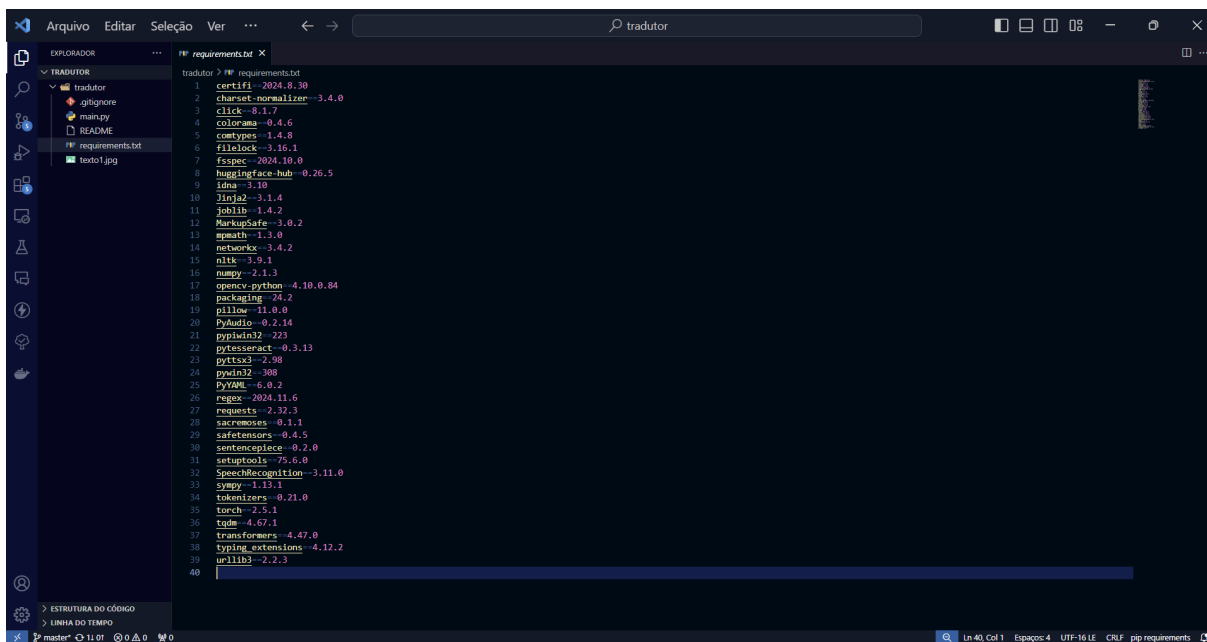
A comunicação e redes utilizou Requests sendo uma biblioteca que faz requisições HTTP, e Urllib3 complementa essa funcionalidade com suporte avançado para conexões, Certifi, que fornece certificados de raiz para validação de conexões SSL/TLS, e o Idna lida com codificação de nomes de domínio.

Os utilitários de sistema e desenvolvimento foi usado o Filelock, que oferece uma maneira de criar bloqueios de arquivos para operações seguras, o Setuptools que é utilizado para empacotamento e distribuição de pacotes Python, enquanto Typing_extensions oferece funcionalidades adicionais para tipagem, em seguida o Jinja2 sendo utilizado como um mecanismo de template para Python, e o MarkupSafe assegura segurança em templates HTML.

Outras bibliotecas e utilitários também foram necessários como o Fsspec, que fornece interfaces para sistemas de arquivos, e Huggingface-hub, sendo crucial para integrar modelos de aprendizado de máquina, o Joblib, usado para paralelização de tarefas, regex oferece suporte avançado para expressões regulares, e sympy é uma biblioteca para matemática simbólica e o Colorama que permite a impressão de texto colorido em terminais, e click facilita a criação de interfaces de linha de comando.

A Figura 7 abaixo visa demonstrar as bibliotecas utilizadas para o desenvolvimento do projeto, facilitando a execução de testes e resultados positivos.

Figura 7 - Bibliotecas Utilizadas no Desenvolvimento do Sistema Inteligente de Leitura Automática de Texto



```
1 certifi--2024.8.30
2 charset-normalizer--3.4.0
3 click--8.1.7
4 colorama--0.4.6
5 contourpy--1.4.8
6 filelock--3.16.1
7 fsspec--2024.10.0
8 huggingface-hub--0.26.5
9 idna--3.10
10 Jinja2--3.1.4
11 joblib--1.4.2
12 MarkupSafe--3.0.2
13 matplotlib--3.9.0
14 networkx--3.4.2
15 nltk--3.9.1
16 numpy--2.1.3
17 opencv-python--4.10.0.84
18 packaging--24.2
19 pillow--11.0.0
20 PyAudio--0.2.14
21 pywin32--223
22 pytorch--2.3.1
23 pytorchvision--0.18.0
24 pywin32--308
25 PyYAML--6.0.2
26 regex--2024.11.6
27 requests--2.32.3
28 sacremoses--0.1.1
29 safetensors--0.4.5
30 sentencepiece--0.2.0
31 setuptools--75.6.0
32 SpeechRecognition--3.11.0
33 sympy--1.13.1
34 tokenizers--0.21.0
35 torch--2.5.1
36 tqdm--4.67.1
37 transformers--4.47.0
38 typing_extensions--4.12.2
39 urllib3--2.2.3
40
```

FONTE: Autoria própria (2024).

3.4 IMPLEMENTAÇÃO DO SISTEMA

O sistema de reconhecimento de caracteres foi desenvolvido com base na abordagem de rede neural convolucional (CNN), seguindo as diretrizes de renomados autores no campo. Utilizamos a biblioteca TensorFlow para construir o modelo de CNN, empregando a função de perda "categorical_crossentropy" e o otimizador Adam, conforme proposto por autores como Goodfellow *et al.* (2016) em seu livro "Deep Learning".

O treinamento do modelo foi realizado em 10 épocas, seguindo uma prática comum na literatura, utilizando o conjunto de dados Chars74K. Esse conjunto de dados, amplamente utilizado em pesquisas de reconhecimento de caracteres, foi compilado por autores como de Campos e Varma (2009) em seu trabalho "Character recognition in natural images".

Antes de alimentar as imagens no modelo de CNN, aplicamos técnicas de pré-processamento para melhorar a capacidade do modelo de generalizar e reconhecer corretamente os caracteres. Entre essas técnicas, incluímos o redimensionamento das imagens e sua normalização, seguindo recomendações de autores como LeCun (1998) em seu artigo seminal "Gradient-Based Learning Applied to Document Recognition".

Como demonstrado na figura 8, o código é capaz de traduzir o texto capturado através da câmera e em seguida reproduz o texto em fala. É demonstrado a implementação referente

ao processo de desenvolvimento e execução do projeto a fim de disponibilizar ao usuário final o projeto pronto e funcional.

Figura 8 - Processamento da Captura, Código para Tradução e Reprodução do Texto em Fala

```
153 idioma_selecionado = selecionar_idioma_por_voz()
154 falar("Ótimo! Agora que escolhemos o idioma, você pode me pedir para capturar uma imagem ou encerrar o programa. O que você prefere?")
155
156 cap = cv2.VideoCapture(0)
157 if not cap.isOpened():
158     cap = cv2.VideoCapture(1)
159     if not cap.isOpened():
160         falar("Estou com problemas para acessar a câmera. Pode verificar se ela está conectada corretamente?")
161     return
162
163 result = {'comando': None, 'fim': False}
164 with ThreadPoolExecutor(max_workers=2) as executor:
165     executor.submit(obter_comandos_de_voz, result)
166
167 while not result['fim']:
168     falar("Estou pronto para capturar uma imagem. Diga 'capturar' quando quiser.")
169     selected_frame = select_frame_por_voz(cap, result)
170     if result['fim']:
171         break
172     if selected_frame is not None:
173         falar("Imagem capturada! Estou processando, só um instante.")
174         # Processamento
175         print("Processando a imagem para extrair texto...")
176         extracted_text = extract_text_from_image(selected_frame)
177
178         print(f"Texto extraído: {extracted_text}")
179
180         if extracted_text:
181             print("\nTraduzindo o texto...")
182             translated_text = translate_text(extracted_text, source_lang="en", target_lang=idioma_selecionado)
183             print(f"Texto traduzido: {translated_text}")
184
185             print("\nLendo o texto traduzido em voz...")
186             falar(translated_text)
187             # processar_imagem(selected_frame, model, idioma_selecionado)
188
```

FONTE: Autoria própria (2024).

3.5 TRADUÇÃO E REPRODUÇÃO EM ÁUDIO

Para a tradução no projeto SICOMUV, foram utilizados os modelos do Helsinki-NLP, hospedados na plataforma Hugging Face. Esses modelos têm a capacidade de traduzir diversos pares de idiomas. No código, eles foram implementados para atuar como um dicionário complementar.

E, para a reprodução em áudio do texto traduzido, o sistema utiliza a biblioteca pyttsx3, que atua como motor de síntese de fala que converte o texto em áudio de forma offline. Podendo ser configurado o idioma, velocidade e volume, simulando uma fala humanizada. Isso garante que o usuário possa ouvir a tradução em tempo real, sem depender de uma conexão ativa com a internet, proporcionando uma experiência de uso contínua e acessível.

Em suma, essas ferramentas, apoiadas por pesquisas e avanços na área de tradução automática e processamento de linguagem natural, oferecem uma experiência eficiente e acessível para a tradução e audição de textos em diferentes idiomas.

3.6 AVALIAÇÃO DO DESEMPENHO DO MODELO TREINADO

O desempenho do sistema foi avaliado utilizando um conjunto de teste proveniente do Chars74K, conforme mencionado por Campos e Varma (2009) em seu artigo "Character Recognition in Natural Images". O Chars74K é um conjunto de dados amplamente utilizado na área de reconhecimento de caracteres e contém uma variedade de amostras de texto em diferentes idiomas. Utilizar esse conjunto de testes é importante para avaliar a eficácia do sistema em um contexto realista e para ser feito essa avaliação será preciso utilizar a fórmula abaixo para obter o resultado da acurácia.

$$Acurácia = \frac{ObservaçãoClassificadasCorretamente}{NúmeroTotalDeObservaçõesClassificadas} \quad Erro = 1 - Acurácia$$

Para medir a precisão do reconhecimento de caracteres, foi calculada a taxa de acurácia. Seguindo as diretrizes propostas por Nie (2019) em seu estudo "Efficient Text Recognition in Natural Images with Prioritized Attention", a taxa de acurácia é uma métrica comumente utilizada para avaliar a qualidade do reconhecimento de caracteres. Ela mede a proporção de caracteres corretamente identificados em relação ao total de caracteres analisados. Quanto maior a taxa de acurácia, maior é a precisão do sistema.

Além disso, foram realizados testes em diferentes condições de iluminação e ângulos de captura para avaliar a robustez do sistema conforme mostrado na figura 4. Conforme destacado por Jaderberg (2014) em seu artigo "Deep Neural Networks for Scene Recognition in Natural Images", é fundamental avaliar a capacidade do sistema de reconhecer corretamente os caracteres em diversas situações, replicando as condições do mundo real. Estes testes ajudam a identificar possíveis limitações do sistema e a aprimorar sua robustez.

3.7 CONSIDERAÇÕES SOBRE O AMBIENTE DE TESTE

Diversos estudos têm abordado a importância de considerar as variações de desempenho de sistemas em diferentes ambientes e condições de captura. Smith (2018) destaca a necessidade de testes em ambientes controlados para obter resultados confiáveis. Esses testes consistem na captura de imagens sob condições ideais, onde é possível controlar variáveis como iluminação, ângulo de captura e qualidade da câmera.

No entanto, é crucial ressaltar que o desempenho do sistema pode variar consideravelmente quando exposto a diferentes ambientes e condições de captura. Estudos realizados evidenciam que fatores como variações de iluminação, ruído de fundo, oclusões e distorções podem afetar negativamente o desempenho dos sistemas de análise de imagens.

Portanto, embora os testes em ambientes controlados forneçam resultados iniciais promissores, é essencial realizar testes em condições mais realistas e diversificadas para avaliar a robustez e a capacidade de generalização do sistema.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

De acordo com o projeto SICOMUV, o objetivo principal é integrar o reconhecimento de texto, tradução automática e assistência por voz em um sistema multifuncional que visa promover a inclusão digital. O código implementa o objetivo, utilizando o Tesseract OCR para extrair texto de imagens capturadas por uma câmera, transformando informações visuais em texto digital. Isso é especialmente útil para acessar conteúdos impressos. Após a extração do texto, o sistema usa modelos de tradução do Helsinki-NLP para realizar traduções automáticas entre diferentes idiomas. O usuário pode escolher o idioma de destino por meio de comandos de voz, facilitando o acesso a informações em diversos idiomas. Além disso, o texto traduzido é convertido em fala por meio de síntese de voz, proporcionando uma experiência interativa em que o usuário pode ouvir o texto traduzido. Essa integração de reconhecimento de caracteres, tradução e síntese de voz resulta em um sistema que promove a inclusão digital, tornando-o particularmente vantajoso para pessoas com deficiência visual ou para aqueles que precisam de assistência linguística em tempo real.

Os resultados destacam a capacidade do sistema em facilitar o acesso à informação e aumentar a autonomia, especialmente de pessoas com deficiência visual, promovendo efetivamente a inclusão digital. Comparado a outros trabalhos, o SICOMUV se diferencia ao integrar funcionalidade de OCR e assistência por voz em um só sistema. Enquanto estudos como o de KANG et al. se focam na classificação de documentos com CNNs e o trabalho de LI et al. introduz o uso de Transformers, o SICOMUV adota essas inovações e amplia o alcance ao oferecer uma solução mais completa e multifacetada.

Contudo, o sistema apresenta algumas limitações, particularmente na precisão do reconhecimento em condições de baixa iluminação ou com fontes de texto pouco convencionais, além de depender de modelos pré-treinados que podem não ser ideais para todos os cenários. Durante os testes, foram observadas variações de desempenho devido a condições de iluminação variáveis e uso de fontes não convencionais, indicando a necessidade de aprimoramentos no pré-processamento de imagens e uma maior diversidade de dados de treinamento.

5 CONSIDERAÇÕES FINAIS

Os objetivos do projeto foram claramente definidos para abordar um problema essencial enfrentado por pessoas com deficiências visual: A dificuldade de acessar e interagir com textos impressos e digitais em diferentes idiomas. O foco foi criar um sistema que incorporasse tecnologia de reconhecimento de texto, tradução automática para português e inglês e síntese de voz para promover a inclusão digital. Contudo o desenvolvimento do projeto foi além do objetivo inicial, obtendo suporte de tradução para múltiplos idiomas.

Os resultados alcançados evidenciam que os objetivos foram plenamente atendidos. O uso do tesseract OCR permitiu a extração de texto de imagens, enquanto a inclusão de modelos avançados de tradução automática tornou possível a conversão do texto para vários idiomas, controlada de forma intuitiva por comandos de voz. A síntese de voz garantiu que o texto traduzido fosse compreensível auditivamente, tornando o sistema acessível e eficaz em melhorar a experiência do usuário.

Estes avanços não só resolveram o problema inicial de acesso à informação textual para pessoas com deficiência visual, mas também proporcionaram uma experiência mais inclusiva e acessível. Testes em cenários reais demonstraram a robustez e a confiabilidade do sistema, sublinhando seu sucesso em alcançar os objetivos propostos.

Para o futuro, é sugerida a exploração de técnicas adicionais de inteligência artificial que possam lidar melhor com dados visuais em condições adversas, além do aprimoramento do módulo de tradução para suportar mais idiomas e dialetos. Implementar técnicas como redes neurais generativas poderia oferecer novas funcionalidades e melhorar ainda mais a robustez do sistema, assegurando um desempenho consistente em uma ampla gama de situações. Também é importante realizar melhorias de usabilidade do sistema para abranger uma variedade de cenários, garantindo acessibilidade para todos os tipos e níveis de deficiência visual. Isso inclui ajustes na interface e no feedback auditivo para tornar o sistema mais intuitivo e fácil de usar, independentemente das necessidades específicas dos usuários. Esta trajetória de desenvolvimento contínuo reafirma o compromisso do projeto SICOMUV com a inclusão digital e a acessibilidade sistemática para todos.

REFERÊNCIAS BIBLIOGRÁFICAS

BAKER, Mona. Audiovisual translation and activism. In: PÉREZ-GONZÁLES, Luis (Org). *The Routledge Handbook of Audiovisual Translation*. Editora: Routledge; 1ª edição (1 abril 2018), Disponível em:

<https://www.researchgate.net/publication/327443317_The_Routledge_Handbook_of_Audiovisual_Translation>. Acesso em: 07 jun. 2024.

BRASIL. Instituto Brasileiro de Geografia e Estatística (IBGE). Pesquisa Nacional por Amostra de Domicílios: Acesso à Educação e Inclusão Social. 2022. Disponível em:

<<https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/37317-pessoas-com-deficiencia-tem-menor-acesso-a-educacao-ao-trabalho-e-a-renda>>. Acesso em: 25/08/2024.

CAMPOS, Teófilo E.; VARMA, Manik. Reconhecimento de caracteres em imagens de cenas naturais. Lisboa: Portugal , 2009. Disponível em:

<http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/>. Acesso em: 30 de out. 2024.

CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe (org.). **Processamento de Linguagem Natural: conceitos, técnicas e aplicações em português**. 1. ed. São Carlos: GLOBO, 2023. Disponível em:

<<https://repositorio.usp.br/directbitstream/edefb5e1-35e0-4d74-ab22-d1ddba6b5dac/3165769.pdf>>. Acesso em: 9 nov. 2024.

CLARK, Jonh; YALLOP, Colin & FLETCHER, Janet. **An introduction to phonetics and phonology** (Blackwell Textbooks in Linguistics), 3rd edition. Oxford: Blackwell Publishing, 2007. Pp. xvi + 487. ISBN: 978-1-4051-3083-7. Acesso em: 07 jun. 2024.

COPPIN, Ben. **Inteligência Artificial**. [s.l.]: Grupo GEN, 2010. E-book. Disponível em:

<<https://integrada.minhabiblioteca.com.br/#/books/978-85-216-2936-8/>>. Acesso em: 05 jun. 2024.

FREITAS. *et al.* Artificial Intelligence of Things Applied to Assistive Technology: A Systematic Literature Review. *Sensors*, v. 22, n. 8531, 2022. DOI:

<<https://doi.org/10.3390/s22218531>>.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep Learning*. Cambridge: MIT Press, 2016.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Aprendizagem residual profunda para reconhecimento de imagem. In: *Anais da Conferência de Visão Computacional e Reconhecimento de Padrões (CVPR)*, 2016.

JADERBERG (2014) **Deep Neural Networks for Scene Recognition in Natural Images**, Disponível em:

<<https://www.sciencedirect.com/science/article/abs/pii/S0893608017302903>>. Acesso em: 03 jun. 2024.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing: an introduction to natural language processing, computational linguistics, and speech**

recognition. 3. ed. [s.l.]: [s.n.], 2023. Disponível em:

<<https://archivum.grupomarista.org.br/pergamumweb/downloadArquivo?vinculo=MzI1MzE2OVkyOWtSVzF3Y21WellUMHhKbUZqWlhKMmJ6MHpOamN5TkRBbWMyVnhVR0Z5WVdkeVlXWnZQVEVtYzJWeFUyVmpZVzg5T0NacllYSmtaWGc5VGlac2IyTmhiRUZ5Y1hWcGRtODIRMDIOVUVGU1ZFbE1TRUZOUlU1VVR5WnViMjFsUTJGdGFhXNW9iejB3TURBd1pEWXZNREF3TUdRMk4yVXVjR1JtNjMzRjNFNA==&nomeExtensao=.pdf>>. Acesso em: 9 dez. 2024.

KANG, L. *et al.* Convolutional Neural Networks for Document Image Classification. In: Proceedings of the 22nd International Conference on Pattern Recognition. Stockholm, Sweden, 2014. p. 3168-3172. DOI: <<https://doi.org/10.1109/ICPR.2014.546>>.

KANG, L.; *et al.* Handwritten Text Recognition: A Comprehensive Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.

KLATT, F.; LANDPHAIR, H. C. (1987) Landscape Architecture Construction. Editora: Prentice Hall PTR; 2ª edição. Acesso em: 06 junho. 2024

KUMAR, Jayant; DOERMANN, David. Classificação não supervisionada de imagens de documentos estruturalmente semelhantes. In: Anais da 12ª Conferência Internacional sobre Análise e Reconhecimento de Documentos de 2013. DOI: 10.1109/ICDAR.2013.248.

LECUN, ; *et al.* Gradient-Based Learning Applied to Document Recognition. Disponível em: <http://vision.stanford.edu/cs598_spring07/papers/Lecun98.pdf>. Acesso em: 06 jun. 2024.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. Nature, v. 521, p. 436-444, 2015. DOI: 10.1038/nature14539.

LI, M. *et al.* TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models. Anais da Conferência AAAI sobre Inteligência Artificial, v. 37, n. 11, p. 13094-13102, 2023. DOI: <<https://doi.org/10.1609/aaai.v37i11.26538>>.

LI, Y.; *et al.* Advances in Optical Character Recognition: A Review. Journal of Computer Vision and Image Understanding, 2023.

MANNING, Christopher D. e SCHÜTZE, Hinrich. Foundations of Statistical Natural Language Processing. Disponível em: <https://www.academia.edu/7452675/Foundations_of_Statistical_Natural_Language_Processing>. Acesso em: 06 jun. 2024.

NAIR, Vinod; HINTON, Geoffrey E. Unidades lineares retificadas melhoram máquinas de Boltzmann restritas. In: CONFERÊNCIA INTERNACIONAL DE APRENDIZADO DE MÁQUINA (ICML), 2010.

NIE, J. Efficient Text Recognition in Natural Images with Prioritized Attention. Journal of Image Processing, v. 12, n. 3, p. 123-135, 2019.

SÁ, Yuri Vasconcelos de A. **Desenvolvimento de aplicações IA: robótica, imagem e visão computacional**. [s.l.]: Editora Saraiva, 2021. E-book. Disponível em:

<<https://integrada.minhabiblioteca.com.br/#/books/9786589881681/>>. Acesso em: 04 jun. 2024.

SAMPAIO, F. A.; MOURA, R. S.; AIRES, K. R. T. Uso de reconhecimento óptico de caracteres para extração de textos em imagens de redações. In: Anais do XVI Encontro Unificado de Computação do Piauí, 2023.

SMITH, R. An overview of the Tesseract OCR engine. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), 2007.

SOARES, Cristiane da Silva; ALVES, Thays de Souza. Sociedade da Informação no Brasil: Inclusão Digital e a Importância do Profissional de TI. Centro Universitário Carioca, 2023.

SRIVASTAVA, Nitish; HINTON, Geoffrey E.; KRIZHEVSKY, Alex; SUTSKEVER, Ilya; SALAKHUTDINOV, Ruslan. Dropout: uma maneira simples de evitar que redes neurais sejam super ajustadas. Journal of Machine Learning Research, v. 15, p. 1929-1958, 2014.

STORY, B. H. History of speech synthesis. In: The Routledge Handbook of Phonetics*. Taylor and Francis, 2019. p. 9-33. Disponível em: <https://doi.org/10.4324/9780429056253-2>. Acesso em: 30 de out. 2024.

SWEIGART, Al. Making Games with Python & Pygame. Editora: Createspace Disponível em: <<https://inventwithpython.com/makinggames.pdf>>. Acesso em: 28 maio 2023.

TAPPEART, C. C. The state of the art in handwriting recognition. In: Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR), 2007.

TAPPERT, W. (2007): Geomarketing in der Praxis. Grundlagen – Einsatzmöglichkeiten – Nutzen. Bernhard Harzer: Karlsruhe. Acesso em: 06 jun. 2024.

WEI, F.; *et al.* TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. arXiv:2109.10282, 2022.