

Sistema de Reconhecimento do Alfabeto da Língua Brasileira de Sinais Utilizando Visão Computacional

Brazilian Sign Language Alphabet Recognition System Using Computer Vision

Aluno: Gabriel Renato Schons, Orientador: Marcos André Lucas¹

¹ Universidade Regional Integrada do Alto Uruguai e das Missões

Departamento de Engenharias e Ciência da Computação

Caixa Postal 743 – 99.709-910 – Erechim – RS – Brasil

101324@aluno.uricer.edu.br, mlucas@uricer.edu.br

Abstract. *This paper presents the development of a computer system capable of recognizing the static letters of the alphabet in Libras (Brazilian Sign Language), using computer vision and machine learning techniques. The goal is to offer an accessible, lightweight, and functional solution for educational and assistive use, capable of operating in real time from video capture. The methodology employed Python, OpenCV, and MediaPipe Hands for extracting the three-dimensional landmarks of the hand, while the Random Forest algorithm was used for gesture classification. Only static signs were considered, excluding letters that involve movement, such as H, J, K, X, Y and Z. The implemented system provides two main functionalities: writing mode, which converts the identified gestures into text, and quiz mode, which assists in learning the letters through practice and feedback from the application. The results obtained demonstrated good performance in different usage conditions, with stability in sign detection and fluid operation in real time. The solution presented contributes to the advancement of assistive technologies aimed at the deaf community, showing potential for both personal use and application in educational contexts.*

Keywords *Sign language, computer vision, gesture classification, MediaPipe, machine learning.*

Resumo. *O presente trabalho apresenta o desenvolvimento de um sistema computacional capaz de reconhecer as letras estáticas do alfabeto em Libras, utilizando técnicas de visão computacional e aprendizado de máquina. O objetivo é oferecer uma solução acessível, leve e funcional para uso educacional e assistivo, capaz de operar em tempo real a partir da captura de vídeo. A metodologia empregou Python, OpenCV e MediaPipe Hands para extração dos landmarks tridimensionais da mão, enquanto o algoritmo Random Forest foi utilizado para a classificação dos gestos. Apenas sinais estáticos foram considerados, excluindo-se letras que envolvem movimento, como H, J, K, X, Y e Z. O sistema implementado disponibiliza duas funcionalidades principais: modo de escrita, que converte os gestos identificados em texto, e modo quiz, que auxilia no aprendizado das letras por meio de prática e feedback da aplicação. Os resultados obtidos demonstraram bom desempenho em diferentes condições de uso, com estabilidade na detecção dos sinais e funcionamento fluido em tempo real. A solução apresentada contribui para o avanço das tecnologias assistivas voltadas à comunidade surda, apresentando potencial tanto para uso pessoal quanto para aplicação em contextos educacionais.*

Palavras-Chave *Libras, visão computacional, classificação de gestos, MediaPipe, aprendizado de máquina.*

1. Introdução

Este trabalho desenvolve um sistema computacional para reconhecimento das letras estáticas do alfabeto em Libras. A solução identifica a mão do usuário via webcam e classifica o gesto correspondente utilizando um modelo treinado com dataset próprio. Foram consideradas apenas letras estáticas, excluindo-se sinais que envolvem movimento para garantir maior consistência e precisão.

A comunicação é um elemento fundamental para a participação social e para o exercício pleno da cidadania. No caso das pessoas surdas, a Língua Brasileira de Sinais (Libras), reconhecida oficialmente pela Lei nº 10.436/2002, constitui o principal meio de expressão e interação. Entretanto, a barreira comunicativa entre surdos e ouvintes ainda é significativa, especialmente devido à falta de domínio da Libras pela população em geral.

Pesquisas em educação bilíngue de surdos indicam que, para grande parte da comunidade surda brasileira, a Libras é a primeira língua, enquanto o português escrito é aprendido como segunda língua, com estrutura gramatical e modalidade muito diferentes (QUADROS; KARNOPP, 2004; INSTITUTO NACIONAL DE EDUCAÇÃO DE SURDOS, 2023). Essa assimetria entre uma língua visuo-espacial e uma língua oral-auditiva escrita torna o processo de alfabetização em português mais complexo e aumenta barreiras comunicativas entre surdos e ouvintes.

Nesse contexto, tecnologias assistivas desempenham papel essencial na inclusão, oferecendo recursos que ampliam a autonomia das pessoas surdas. Avanços recentes em visão computacional e aprendizado de máquina possibilitaram o desenvolvimento de ferramentas capazes de interpretar gestos em tempo real, reduzindo a necessidade de equipamentos específicos ou ambientes controlados. Segundo Szeliski (2022), a visão computacional é a área da inteligência artificial que desenvolve algoritmos capazes de extrair, analisar e compreender informações visuais a partir de imagens e vídeos, permitindo que sistemas computacionais realizem tarefas como detecção, reconhecimento e rastreamento de objetos. Essa característica torna o campo especialmente adequado para aplicações de reconhecimento gestual e interpretação automática de línguas de sinais.

O sistema desenvolvido oferece duas funcionalidades adicionais que ampliam sua utilidade prática e pedagógica. A primeira delas é o modo escrita, no qual os gestos reconhecidos pela câmera são automaticamente convertidos em caracteres exibidos na tela. Esse recurso possibilita que o usuário produza palavras ou pequenas frases utilizando exclusivamente sinais da Libras, funcionando tanto como meio de comunicação quanto como ferramenta auxiliar para estudantes em processo de alfabetização visual-manual. Já a segunda funcionalidade corresponde ao modo quiz, que apresenta desafios interativos com letras aleatórias do alfabeto em Libras e estimula o usuário a reproduzir corretamente cada gesto. Esse modo promove o aprendizado ativo, incentiva a memorização dos sinais e torna o estudo mais dinâmico, servindo também como apoio a práticas educativas em sala de aula ou ambientes informais. Juntas, essas funcionalidades ampliam o potencial didático e inclusivo do sistema, reforçando sua relevância tanto no contexto comunicativo quanto no pedagógico.

De maneira geral, os principais objetivos deste trabalho concentram-se no desenvolvimento de um sistema capaz de realizar o reconhecimento, em tempo real, das letras estáticas do alfabeto em Libras, além da criação de uma ferramenta acessível que contribua para a inclusão comunicativa e auxilie o processo de ensino e aprendizagem da língua. A proposta visa proporcionar uma solução leve, funcional e baseada em tecnologias modernas, permitindo que seja executada facilmente em computadores convencionais e utilizada por diferentes perfis de usuários, como

estudantes, professores, familiares de pessoas surdas e qualquer pessoa interessada em aprender Libras. O projeto também busca incentivar o desenvolvimento de futuras pesquisas, abrindo caminho para o reconhecimento de gestos dinâmicos, sinais complexos e até estruturas linguísticas completas.

Ao final desta introdução, os demais capítulos estruturam-se da seguinte forma: os Trabalhos Relacionados e a Fundamentação Teórica apresentam os conceitos e tecnologias que embasam o projeto; a Metodologia detalha os procedimentos adotados; a Implementação descreve o desenvolvimento do sistema; os Resultados e Discussão analisam o desempenho do modelo; e a Conclusão e Trabalhos Futuros apresentam considerações finais, limitações e possibilidades de aprimoramento, destacando o potencial expansível da solução construída.

2. Trabalhos Relacionados e Fundamentação Teórica

O reconhecimento automático de sinais da Língua Brasileira de Sinais (Libras) tem sido tema recorrente em pesquisas nas áreas de visão computacional, tecnologia assistiva e aprendizado de máquina. Segundo Goodfellow, Bengio e Courville (2016), sistemas de percepção computacional têm avançado rapidamente devido à capacidade dos modelos modernos de aprender representações complexas diretamente a partir dos dados. No contexto de Libras, tais tecnologias contribuem diretamente para a inclusão comunicativa, ampliando o acesso de pessoas surdas a ferramentas digitais.

A evolução da área revela uma diversidade de abordagens, ferramentas e arquiteturas que buscam interpretar gestos manuais com maior precisão, eficiência e aplicabilidade em ambientes reais. Este capítulo apresenta uma revisão das principais categorias de soluções existentes, destacando suas características técnicas, limitações e como o presente trabalho se posiciona em relação a elas, assim como a fundamentação teórica que embasou este trabalho.

2.1. Abordagens Existentes na Literatura

As técnicas baseadas em imagens tradicionais foram amplamente utilizadas em estudos iniciais, empregando segmentação de mão, limiarização e extração manual de características. Para Viola e Jones (2001), métodos clássicos dependem fortemente de condições controladas para manter sua estabilidade, o que limita sua aplicação prática. Embora tais abordagens tenham demonstrado viabilidade, apresentam forte dependência de iluminação adequada e fundos uniformes.

Com os avanços do aprendizado profundo, as soluções passaram a adotar redes neurais convolucionais (CNNs) para reconhecimento gestual. LeCun, um dos pioneiros das CNNs, afirma que essas redes são capazes de extrair automaticamente padrões altamente discriminativos, alcançando acurácia elevada em tarefas visuais complexas (LECUN; BENGIO; HINTON, 2015). No entanto, tais métodos exigem grandes datasets, alto poder computacional e frequentemente GPU, o que limita sua adoção em dispositivos comuns — problema já apontado também por Krizhevsky, Sutskever e Hinton (2012) em grandes modelos visuais.

Mais recentemente, tornaram-se populares os métodos baseados em landmarks, que utilizam modelos otimizados para detectar pontos anatômicos tridimensionais da mão. Holzbauer e Zhang (2020) destacam que esses métodos apresentam excelente estabilidade e independência das condições de luz. Esta família de técnicas foi impulsionada principalmente após a publicação do MediaPipe Hands por Lugaresi et al. (2019), que demonstraram a viabilidade de detecção de 21 pontos da mão com baixo custo computacional em tempo real.

2.2. Lacunas Identificadas

A análise das abordagens existentes revela um conjunto de lacunas relevantes que ainda persistem na literatura e que motivam o desenvolvimento de soluções mais acessíveis e adequadas ao contexto brasileiro. Um dos primeiros pontos observados diz respeito à forte dependência de ambientes controlados, característica marcante dos métodos tradicionais de detecção de padrões visuais, como o modelo proposto por Viola e Jones (2001). Esses métodos exigem condições específicas de iluminação, posicionamento e contraste para garantir resultados satisfatórios, o que limita sua aplicação em cenários reais e reduz a usabilidade para usuários comuns. Esse tipo de restrição torna o reconhecimento gestual pouco adaptável a ambientes educativos ou domésticos, onde fatores externos variam constantemente.

Além disso, os métodos modernos baseados em redes neurais convolucionais (CNN) — como os apresentados por Krizhevsky, Sutskever e Hinton (2012) —, embora altamente precisos, demandam hardware potente, geralmente com suporte a GPUs, para alcançar desempenho adequado. Essa exigência tecnológica é um obstáculo significativo em contextos de escolas públicas, instituições de ensino com recursos limitados ou usuários que dispõem apenas de computadores convencionais. Assim, apesar do avanço representado pelas CNNs, sua adoção ampla ainda é limitada pelo custo computacional e pelas dificuldades de implementação.

Outra lacuna importante refere-se à escassez de soluções voltadas especificamente ao contexto educacional e assistivo. Grande parte das tecnologias desenvolvidas para surdos concentra-se na tradução automática ou na interpretação de sinais complexos, mas poucas iniciativas são destinadas ao ensino do alfabeto, ao apoio didático ou à alfabetização em Libras. Há, portanto, uma falta de ferramentas simples, intuitivas e direcionadas tanto para iniciantes quanto para ambientes pedagógicos, o que reduz o acesso a recursos de aprendizado acessíveis e gratuitos.

Por fim, observa-se a carência de sistemas leves capazes de operar em tempo real mesmo em hardware simples. A maioria dos modelos existentes prioriza desempenho e alta precisão em detrimento da praticidade, resultando em soluções que, embora tecnicamente avançadas, não são viáveis para uso cotidiano. Essa limitação reforça a necessidade de ferramentas mais inclusivas, que funcionem com eficiência em computadores de baixo custo e que não dependam de grandes estruturas computacionais.

Considerando esses aspectos, torna-se evidente que ainda há um espaço significativo para o desenvolvimento de sistemas mais acessíveis, eficientes e alinhados às demandas reais do ensino e da prática da Libras. É nesse conjunto de lacunas que este trabalho se insere, propondo uma solução que busca equilibrar simplicidade, desempenho e utilidade pedagógica.

A Tabela 1 apresenta a comparação entre a principal abordagem de reconhecimento gestual, utilizado em sistemas já existentes, e a implementação de reconhecimento a partir de landmarks.

Tabela 1. Comparação entre abordagens de reconhecimento gestual

Critério	Deep Learning (CNN)	Landmarks
Sensibilidade à iluminação	Alta	Baixa
Dependência de fundo uniforme	Média	Nenhuma
Necessidade de pré-processamento	Baixa	Baixa
Exigência de hardware	Alta	Baixa
Acurácia	Alta	Alta
Execução em tempo real	Média	Alta
Complexidade	Alta	Baixa

2.3. Posicionamento do Trabalho Atual

O sistema desenvolvido nesta pesquisa se apoia na abordagem baseada em landmarks, combinando robustez, eficiência e acessibilidade. De acordo com Lugaresi (2019), o uso de pontos anatômicos tridimensionais permite operar independentemente da iluminação ou do ambiente, garantindo estabilidade mesmo em cenários não controlados.

Além disso, o uso de um classificador leve, como o Random Forest, segue recomendações de Breiman (2001), que destaca a capacidade desse algoritmo de obter excelente acurácia mesmo com baixo custo computacional, reforçando sua adequação para aplicações em tempo real.

O sistema diferencia-se das soluções observadas na literatura ao integrar reconhecimento gestual com funcionalidades interativas, ampliando seu potencial educacional. Tecnologias assistivas realmente eficazes devem priorizar acessibilidade, baixo custo e aplicabilidade pedagógica, ou seja, exatamente o foco deste trabalho.

Assim, o trabalho atual se posiciona como uma solução moderna, prática e acessível, preenchendo lacunas relacionadas à leveza computacional, interatividade e aplicabilidade pedagógica, contribuindo diretamente para a inclusão comunicativa por meio da Libras.

Para compreender o funcionamento e a importância de um sistema de reconhecimento do alfabeto da Libras por meio de visão computacional e aprendizado de máquina, é necessário explorar de forma aprofundada temas como: (i) a Língua Brasileira de Sinais e suas características linguísticas; (ii) acessibilidade e tecnologias assistivas; (iii) visão computacional e reconhecimento de gestos; (iv) o funcionamento do MediaPipe Hands; (v) conceitos de aprendizado de máquina e classificação; e (vi) estudos relacionados. Cada um desses tópicos é apresentado a seguir, compondo um panorama teórico robusto que embasa o projeto desenvolvido.

2.4. Língua Brasileira de Sinais (Libras)

A Língua Brasileira de Sinais (Libras) é uma língua natural, visual-motora, dotada de estrutura gramatical própria e reconhecida como meio legal de comunicação desde a sanção da Lei nº 10.436, de 24 de abril de 2002. Essa legislação, complementada pelo Decreto nº 5.626/2005, não apenas reconhece a Libras, mas estabelece diretrizes para sua difusão e para a formação de profissionais aptos a utilizá-la.

Do ponto de vista linguístico, a Libras é descrita como uma língua natural completa, com níveis fonológico, morfológico, sintático e discursivo organizados na modalidade visuo-espacial, o que demanda descrições gramaticais próprias e independentes do português (QUADROS; KARNOPP, 2004; QUADROS, 2009). Esses estudos consolidam a Libras como objeto legítimo de investigação linguística e fundamentam a importância de recursos tecnológicos que respeitem suas especificidades estruturais.

Dessa forma, a Libras constitui elemento essencial da identidade cultural e da autonomia comunicativa da comunidade surda no Brasil. Sistemas automatizados que reconhecem sinais da Libras têm potencial para expandir o acesso à comunicação e reduzir barreiras que ainda persistem no convívio social.

A Figura 1 apresenta o alfabeto manual da Língua Brasileira de Sinais, no qual cada letra é representada por uma configuração específica da mão. A maioria dos sinais é estática, formada apenas pela posição dos dedos, enquanto algumas letras requerem movimento, indicado pelas setas na figura.

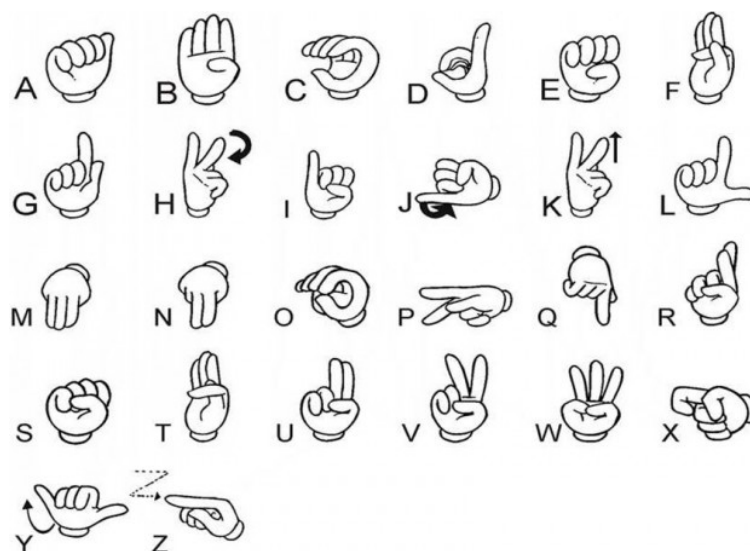


Figura 1. Alfabeto em Libras

2.5. Acessibilidade e Tecnologias Assistivas

A acessibilidade comunicacional é um direito garantido pela Lei Brasileira de Inclusão da Pessoa com Deficiência (Lei nº 13.146/2015), que determina a adoção de tecnologias assistivas como forma de remover barreiras e ampliar a participação social. Estima-se que cerca de 15% da população mundial viva com algum tipo de deficiência, conforme apontado pela Organização Mundial da Saúde e pelo Banco Mundial, o que torna a discussão sobre acessibilidade um tema central nas políticas públicas e nas práticas sociais voltadas à inclusão (WORLD HEALTH ORGANIZATION; WORLD BANK, 2011). Esses dados evidenciam a necessidade de iniciativas que promovam autonomia e igualdade de oportunidades, sobretudo em contextos comunicativos, onde barreiras linguísticas podem impactar diretamente o acesso à educação, ao trabalho e aos serviços básicos.

O Relatório Mundial sobre a Deficiência, publicado pela Organização Mundial da Saúde em parceria com o Banco Mundial, define tecnologia assistiva como o conjunto de produtos, equipamentos, dispositivos, recursos, metodologias e serviços que têm como finalidade ampliar a funcionalidade, a autonomia e a participação social de pessoas com deficiência. Essa definição abrange desde equipamentos físicos até soluções digitais capazes de facilitar a comunicação e a interação em diferentes ambientes (WORLD HEALTH ORGANIZATION; WORLD BANK, 2011). No cenário brasileiro, a Lei Brasileira de Inclusão reforça esse entendimento ao estabelecer que o acesso à tecnologia assistiva é um direito fundamental e um mecanismo indispensável para a eliminação de barreiras, a promoção da acessibilidade e o fortalecimento da inclusão social (BRASIL, 2015).

No contexto específico da surdez, tecnologias assistivas desempenham papel ainda mais relevante ao atuar como mediadoras entre diferentes modalidades linguísticas. Entre os recursos mais utilizados estão os aplicativos de tradução de textos para Libras, que permitem converter conteúdos escritos para representações visuais acessíveis; os intérpretes automáticos, que buscam transpor a barreira entre a língua oral e a língua de sinais; os softwares educativos, voltados ao ensino da Libras e ao apoio pedagógico; e os sistemas de reconhecimento de gestos, que utilizam visão computacional e inteligência artificial para identificar sinais realizados manualmente. Esses recursos, embora distintos em suas funcionalidades, compartilham o objetivo comum de facilitar a comunicação e reduzir desigualdades linguísticas enfrentadas diariamente pela comunidade surda.

A adoção de sistemas computacionais capazes de automatizar a interpretação de sinais da Libras, especialmente aqueles que funcionam em tempo real, representa um avanço significativo para a promoção da inclusão. Esses sistemas podem atuar como ponte comunicativa entre surdos e ouvintes, permitindo interações mais naturais e menos dependentes da presença de intérpretes humanos. Além disso, eles funcionam como ferramentas de apoio educacional, auxiliando no aprendizado da língua de sinais por estudantes, professores, familiares ou qualquer pessoa interessada em estabelecer uma comunicação mais acessível com a comunidade surda. Dessa forma, a inclusão digital promovida por tais soluções contribui diretamente para ampliar a autonomia das pessoas surdas e fortalecer a construção de uma sociedade mais equitativa e inclusiva.

2.6. Visão Computacional

A visão computacional é uma área da inteligência artificial dedicada ao desenvolvimento de algoritmos capazes de extrair informação útil a partir de imagens e vídeos, aproximando-se, em alguma medida, da capacidade humana de perceber o ambiente visual. Szeliski (2022) descreve o campo como o estudo de técnicas que recebem dados de imagem como entrada e produzem descrições de alto nível da cena, permitindo realizar tarefas como detecção e reconhecimento de objetos, reconstrução tridimensional, análise da geometria do ambiente e rastreamento de movimentos. Essa definição evidencia o caráter multidisciplinar da área, que combina fundamentos de matemática, estatística, processamento de sinais e aprendizado de máquina para compreender padrões visuais complexos com precisão crescente.

O processo básico de visão computacional envolve diversas etapas articuladas entre si. A primeira delas é a aquisição da imagem, geralmente realizada por meio de câmeras convencionais, webcams ou sensores visuais mais avançados. Após a captura, inicia-se o pré-processamento, etapa responsável por preparar a imagem para análises posteriores. Esse procedimento pode incluir ajustes de brilho e contraste, normalização dos níveis de cor, redução de ruído, aplicação de filtros e correções geométricas, tudo com o objetivo de destacar informações relevantes e minimizar interferências do ambiente. Em seguida, ocorre a extração de características, fase na qual o algoritmo identifica elementos distintivos presentes na cena, como bordas, texturas, contornos, regiões de interesse ou pontos específicos que podem representar informações estruturais do objeto analisado. Por fim, essas características são enviadas para um módulo de classificação, que utiliza métodos de aprendizado de máquina ou redes neurais para categorizar os padrões encontrados e fornecer interpretações sobre o conteúdo da imagem.

No reconhecimento de sinais manuais, essas etapas tornam-se ainda mais específicas. O sistema normalmente inicia realizando a segmentação da mão, isolando-a do restante da imagem por meio de técnicas como limiarização, análise de profundidade ou modelos de detecção baseados em aprendizado profundo. Depois disso, ocorre a identificação dos contornos da mão, que permite compreender sua forma geral e distinguir os limites dos dedos. A análise da posição dos dedos é fundamental, pois cada gesto de Libras depende da configuração exata das articulações e da abertura ou fechamento dos dedos. Tecnologias modernas, como MediaPipe Hands, aprofundam esse processo ao extrair pontos-chave tridimensionais (landmarks), representando com precisão a posição de articulações específicas. Essas coordenadas servem como base para algoritmos de classificação que interpretam cada configuração manual como um determinado sinal.

Historicamente, o reconhecimento automático de gestos enfrentou diversas limitações. Trabalhos mais antigos utilizavam segmentação baseada em cor da pele, exigindo condições de iluminação controladas para evitar distorções; em outros casos, dependiam de luvas marcadas

com cores específicas ou sensores adicionais, o que restringia a mobilidade e a naturalidade do gesto. Alguns sistemas também faziam uso de dispositivos caros, como sensores infravermelhos ou câmeras de profundidade, inviabilizando sua adoção em larga escala. Com o avanço da inteligência artificial e a popularização de métodos mais robustos, como redes neurais profundas e modelos de landmarks tridimensionais, tornou-se possível desenvolver soluções mais eficientes, capazes de operar em tempo real e com alto grau de precisão utilizando apenas webcams comuns. Essa evolução democratiza o acesso a ferramentas de visão computacional e possibilita aplicações mais inclusivas e acessíveis, especialmente no reconhecimento de sinais da Libras.

2.7. Reconhecimento de Gestos e Sinais

O reconhecimento de gestos consiste na identificação automática de padrões manuais a partir de imagens ou vídeos, permitindo que sistemas computacionais interpretem movimentos e posições da mão de maneira semelhante ao processamento visual humano. Essa técnica tem se tornado cada vez mais relevante em diversas áreas, sendo amplamente aplicada em jogos que utilizam controle por movimento, em sistemas robóticos que dependem de comandos gestuais para interação intuitiva, em ambientes educacionais que exploram recursos visuais para ensino e aprendizagem, e, sobretudo, em ferramentas dedicadas à tradução e interpretação de línguas de sinais. Sua versatilidade demonstra a importância crescente de métodos capazes de compreender sinais manuais de forma precisa e natural.

No campo do reconhecimento gestual, duas abordagens principais se consolidaram ao longo dos anos. A primeira delas é a abordagem baseada diretamente na imagem, que utiliza dados brutos provenientes de câmeras RGB ou grayscale. Nesse caso, o sistema trabalha com pixels da imagem original, podendo incluir etapas de segmentação, binarização e aplicação de técnicas de visão computacional tradicional ou redes neurais convolucionais (CNNs). Embora essa abordagem seja poderosa e tenha alcançado resultados expressivos, especialmente com o avanço das CNNs, ela costuma demandar maior capacidade computacional e sofrer com variações de iluminação, ângulos e ruídos visuais.

A segunda abordagem, que vem ganhando destaque, é a baseada em landmarks, ou seja, em pontos anatômicos específicos da mão — como articulações, pontas dos dedos e orientações das falanges. Esses pontos são extraídos por modelos especializados, como o MediaPipe Hands, que captura com precisão a posição tridimensional de dezenas de pontos relevantes mesmo em condições adversas. A utilização de landmarks oferece diversas vantagens, pois reduz significativamente a sensibilidade à iluminação, aumenta a velocidade de inferência e diminui a necessidade de pré-processamento, além de fornecer maior consistência entre usuários com diferentes tonalidades de pele, tamanhos de mão ou ambientes de filmagem distintos. A representação da mão por meio de coordenadas estruturadas permite que algoritmos de aprendizado de máquina foquem nos padrões mais importantes da configuração gestual, resultando em classificações mais estáveis.

Estudos recentes sobre o reconhecimento automático de línguas de sinais reforçam a eficácia dessa abordagem. Pesquisas como as de Furtado (2021) e Silva (2021) demonstram que a combinação de descritores baseados na configuração da mão com modelos de aprendizado supervisionado ou redes neurais profundas é capaz de atingir resultados robustos e operar em tempo real, inclusive para sinais da Libras. Esses trabalhos destacam o potencial de soluções que utilizam landmarks e algoritmos inteligentes para interpretar sinais capturados diretamente por webcams, reduzindo a necessidade de dispositivos adicionais ou condições controladas. Assim, a literatura contemporânea corrobora a escolha metodológica adotada neste projeto, que também

explora a extração de pontos-chave da mão como base para o reconhecimento preciso das letras estáticas da Libras.

2.8. O MediaPipe Hands

O MediaPipe Hands é um framework desenvolvido e fundamentado na arquitetura Blaze, desenvolvida por pesquisadores do Google Research, reconhecida internacionalmente pela eficiência na detecção em tempo real. A abordagem proposta por Bazarevsky. (2019) introduz modelos ultra rápidos para detecção de rosto e palma da mão, os quais serviram de base para a implementação posterior do MediaPipe Hands. Essa fundamentação torna o método altamente adequado para aplicações interativas, permitindo calcular 21 landmarks tridimensionais da mão, permitindo reconstruir a postura dos dedos mesmo em condições de iluminação variável ou ângulos não convencionais.

Essa tecnologia revolucionou aplicações educacionais, assistivas e interativas, pois elimina a necessidade de:

- Fundo uniforme;
- Câmera especial;
- Sensores adicionais;
- Luvas marcadas.

Diferentemente de sistemas que dependem de CNNs pesadas para interpretar a imagem bruta, o MediaPipe fornece um vetor numérico direto (21×3 coordenadas), simplificando enormemente o processo de classificação.

2.9. Aprendizado de Máquina e Classificação de Gestos

O aprendizado de máquina é uma subárea central da inteligência artificial dedicada ao estudo de algoritmos capazes de identificar padrões complexos a partir de dados e ajustar automaticamente seus parâmetros de modo a realizar tarefas como classificação, regressão, previsão temporal e detecção de anomalias. Em vez de serem programados explicitamente para cada situação, esses algoritmos aprendem com exemplos, extraindo regularidades estatísticas que permitem realizar inferências sobre novos dados. Segundo Bishop (2006) e Goodfellow, Bengio e Courville (2016), essa capacidade de generalização é fundamental para que modelos se tornem úteis em cenários reais, nos quais existe ruído, variação individual e mudanças no ambiente.

No caso da aprendizagem supervisionada, utilizada neste projeto, o modelo recebe durante o treinamento conjuntos de entrada e saída (features e rótulos) que descrevem o comportamento desejado. A partir dessa estrutura, o algoritmo aprende uma função que aproxima o mapeamento entre os padrões observados e suas respectivas classes, tornando-se capaz de prever corretamente novos exemplos que seguem lógica semelhante. Uma das vantagens dessa abordagem é sua flexibilidade, permitindo que tanto dados visuais quanto vetores numéricos sejam utilizados como entrada, que é característica essencial para problemas envolvendo landmarks da mão.

No contexto do reconhecimento gestual baseado em landmarks, o processo segue uma sequência de etapas essenciais para a construção de um sistema confiável. Primeiramente, são coletados vetores contendo as coordenadas dos pontos anatômicos da mão, obtidos a partir do rastreamento realizado pelo MediaPipe Hands ou mecanismos equivalentes. Esses vetores, compostos por dezenas de coordenadas bidimensionais ou tridimensionais, são então rotulados com a letra correspondente, formando um dataset estruturado. Em seguida, ocorre o treinamento do classificador, que aprende a identificar padrões geométricos específicos de

cada configuração manual. Finalmente, o modelo treinado passa a ser utilizado em tempo real, recebendo continuamente novos vetores e classificando instantaneamente os gestos executados pelo usuário. Esse ciclo de captura, rotulagem, aprendizagem e inferência constitui a base dos sistemas modernos de reconhecimento gestual.

Para realizar essa tarefa, diferentes algoritmos podem ser utilizados, cada um com características próprias. Redes neurais, por exemplo, conseguem modelar relações altamente não lineares e são particularmente eficazes em grandes conjuntos de dados. SVMs (Support Vector Machines) destacam-se quando há necessidade de margens de decisão bem definidas e apresentam excelente desempenho com datasets menores ou de alta dimensionalidade. Já o Random Forest, classificador adotado neste trabalho, apresenta uma combinação de simplicidade, robustez e bom desempenho computacional, especialmente ao lidar com dados tabulares e variações naturais presentes nas coordenadas dos landmarks.

Proposto por Breiman (2001), o Random Forest é composto por uma “floresta” de árvores de decisão independentes, cada uma treinada a partir de subconjuntos aleatórios de amostras e atributos. Essa estratégia de aleatoriedade controlada aumenta a diversidade das árvores, reduzindo significativamente a probabilidade de overfitting. A predição final do modelo é obtida por votação entre as árvores individuais, garantindo maior estabilidade e confiabilidade nos resultados. Além disso, o método é eficiente para lidar com dados ruidosos, variações individuais entre usuários e pequenas imperfeições na captura dos landmarks, o que o torna particularmente adequado para o reconhecimento das letras estáticas da Libras.

Essa combinação de eficiência computacional, facilidade de implementação e robustez diante de dados heterogêneos justifica a adoção do Random Forest no presente trabalho, demonstrando ser uma abordagem compatível com o objetivo de construir um sistema leve, acessível e capaz de operar em tempo real.

3. Metodologia

A metodologia adotada neste trabalho descreve como o estudo foi conduzido, incluindo o tipo de pesquisa, os métodos utilizados, as ferramentas selecionadas e as etapas que compõem o desenvolvimento do sistema de reconhecimento das letras estáticas da Libras. O foco deste capítulo está nas decisões metodológicas, nos procedimentos experimentais e na justificativa das escolhas realizadas.

O presente estudo caracteriza-se como uma pesquisa aplicada, pois busca desenvolver uma solução prática voltada à mediação comunicativa entre surdos e ouvintes por meio do reconhecimento automático de sinais. Trata-se também de uma pesquisa experimental, uma vez que envolve testes controlados para validação do desempenho do sistema, e apresenta caráter quantitativo, devido ao uso de métricas objetivas, como acurácia, matriz de confusão e taxa de acertos por classe.

A escolha das tecnologias adotadas foi fundamentada na necessidade de desenvolver um sistema leve, acessível e capaz de operar em tempo real. As principais ferramentas utilizadas foram:

- Python, como linguagem de alto nível amplamente usada em visão computacional;
- OpenCV, para captura de vídeo e manipulação dos frames;
- MediaPipe Hands, responsável pela detecção dos 21 landmarks da mão;
- NumPy, para manipulação numérica eficiente;
- scikit-learn, utilizada para treinamento do algoritmo de classificação;

- HTML, CSS e JavaScript, empregados na construção da interface web.

A escolha do algoritmo Random Forest baseou-se em sua robustez, eficiência em datasets pequenos e baixa suscetibilidade a ruído, sendo adequado para aplicações em tempo real.

3.1. Etapas do Desenvolvimento e Justificativa

O processo metodológico adotado neste trabalho foi estruturado em etapas sequenciais e iterativas, permitindo a construção gradual do sistema e o aperfeiçoamento contínuo dos resultados. Inicialmente, foi realizada a elaboração do dataset, construído manualmente com o objetivo de contemplar exclusivamente as letras estáticas do alfabeto em Libras. Optou-se por excluir as letras H, J, K, X, Y e Z, pois tais sinais envolvem movimentos contínuos que exigiriam outro tipo de abordagem e aumentariam significativamente a complexidade do modelo. A delimitação desse escopo possibilitou maior consistência na captura dos dados e tornou o treinamento mais adequado ao tipo de algoritmo escolhido.

As imagens utilizadas para compor o dataset foram capturadas pela própria webcam do notebook, com nitidez e detalhes nas estruturas da mão. Durante cada captura, o MediaPipe Hands detectava automaticamente os 21 pontos anatômicos (landmarks) da mão, retornando, para cada um deles, os valores tridimensionais (x, y, z). Esses valores eram armazenados de forma estruturada, e o procedimento inteiro foi repetido centenas de vezes para cada letra, assegurando a formação de um conjunto de dados significativo e diversificado. Cada amostra gerada contém um total de 63 valores numéricos, resultantes da concatenação das coordenadas dos 21 landmarks, formando um vetor padronizado capaz de representar a configuração manual correspondente ao sinal capturado.

Posteriormente, iniciou-se a etapa de normalização dos dados, fundamental para reduzir variações indesejadas entre amostras e garantir maior estabilidade durante o treinamento do modelo. Esse procedimento visou diminuir a influência da distância da mão até a câmera, minimizar diferenças naturais entre usuários — como tamanho da mão, comprimento dos dedos ou amplitude dos gestos — e assegurar a consistência entre os vetores de entrada. Para isso, foram aplicadas técnicas como a divisão dos valores pelo tamanho relativo da mão e o cálculo de proporções entre pontos anatômicos, permitindo que o modelo aprendesse padrões estruturais e não dependentes de dimensões absolutas. Essa estratégia contribuiu significativamente para tornar o sistema mais robusto e adaptável a diferentes condições de uso.

Com os dados preparados, o conjunto foi dividido em duas partes: 80% destinado ao treinamento do modelo e 20% reservado para teste, utilizando estratificação para garantir que todas as classes estivessem proporcionalmente representadas. Essa divisão permitiu avaliar o desempenho do classificador de forma confiável, assegurando que os resultados obtidos refletissem a capacidade real de generalização do modelo. As métricas utilizadas durante a validação incluíram a acurácia geral, a matriz de confusão, a taxa de acertos por classe e o tempo médio de inferência. Cada uma dessas métricas forneceu uma perspectiva complementar: enquanto a acurácia mostrou o desempenho agregado, a matriz de confusão permitiu identificar classes mais suscetíveis a erros, e o tempo de inferência avaliou a adequação do modelo para aplicações em tempo real.

Essas etapas metodológicas foram acompanhadas de justificativas que orientaram a escolha das ferramentas e decisões de implementação. Priorizaram-se critérios como desempenho — visando garantir operação fluida mesmo sem o uso de GPU — e acessibilidade tecnológica, de modo que o sistema pudesse funcionar utilizando apenas uma webcam comum em hardware modesto. A robustez também foi um fator decisivo, especialmente no que se refere à insensibilidade do modelo a variações de iluminação ou fundos não controlados, características

essenciais para viabilizar o uso do sistema em ambientes reais. Além disso, buscou-se simplicidade no processo de captura de dados, uma vez que o uso de landmarks dispensa etapas complexas de pré-processamento e reduz significativamente o custo computacional da tarefa.

O uso de landmarks, portanto, demonstrou-se altamente vantajoso ao oferecer maior consistência na representação das posições anatômicas da mão, ao mesmo tempo em que simplificou o pipeline de processamento e tornou o sistema mais leve, rápido e acessível. Essa escolha metodológica se alinha ao objetivo central do projeto, que é desenvolver uma solução capaz de operar em tempo real de forma eficiente e adequada a ambientes educacionais e ao uso por usuários comuns.

4. Implementação

Este capítulo descreve o processo de desenvolvimento do sistema, apresentando o ambiente utilizado, a arquitetura do código, as decisões técnicas, os desafios encontrados e as estratégias de otimização.

4.1. Ambiente de Desenvolvimento

O desenvolvimento foi realizado no:

- Visual Studio Code, como IDE principal;
- Notebook Dell G3, Intel i5, 16 GB RAM, com webcam HD.

Todo o sistema foi construído com ferramentas multiplataforma e de código aberto, garantindo portabilidade e fácil reprodução por outros usuários.

4.2. Arquitetura do Sistema

A arquitetura do sistema foi organizada em módulos independentes, permitindo uma estrutura clara, escalável e de fácil manutenção. Essa modularização foi planejada para garantir que cada componente desempenhasse uma função específica dentro do fluxo de reconhecimento, facilitando tanto o desenvolvimento quanto futuras expansões. Dessa forma, tornou-se possível separar de maneira nítida a lógica de processamento, a parte visual da aplicação e os algoritmos de inteligência artificial envolvidos.

O processo inicia-se pelo módulo de captura de vídeo, responsável por obter continuamente os frames da webcam por meio da biblioteca OpenCV. Esse módulo é fundamental, pois fornece a base visual sobre a qual o sistema opera, garantindo uma captura estável e em tempo real. Em seguida, cada frame é encaminhado ao módulo de detecção de landmarks, executado pelo MediaPipe Hands, que extrai automaticamente os 21 pontos anatômicos da mão e suas coordenadas tridimensionais. Essa etapa é crucial para transformar a informação visual em dados estruturados, adequados para análise computacional.

O vetor de 63 valores gerado pela detecção é então encaminhado ao módulo de pré-processamento, que realiza a normalização e padronização dos dados. Esse módulo organiza e ajusta as coordenadas para eliminar variações relacionadas ao tamanho da mão, à distância da câmera ou ao posicionamento no quadro, garantindo que todos os vetores sigam um padrão comum. Após essa preparação, os dados são enviados ao módulo de classificação, onde o modelo Random Forest (previamente treinado e carregado na memória) realiza a inferência e determina qual letra está sendo exibida pelo usuário. Essa etapa ocorre em milissegundos, possibilitando respostas rápidas e compatíveis com aplicações interativas.

Por fim, os resultados são encaminhados ao módulo de interface web, responsável por exibir as letras identificadas, registrar a interação do usuário e gerenciar funcionalidades como o Modo Escrita e o Modo Quiz. A interface atua como ponto de contato entre o sistema e o usuário final, apresentando as informações de forma acessível e amigável.

A adoção dessa abordagem modular trouxe benefícios significativos. Além de facilitar a integração entre Python (responsável pelo processamento e pela inteligência artificial) e a interface web (responsável pela exibição e interação), ela permitiu que cada parte fosse desenvolvida, testada e aperfeiçoada separadamente. Essa independência entre os componentes torna o sistema mais flexível, possibilita substituições futuras (como a troca do modelo de classificação ou da tecnologia de captura) e garante maior eficiência no fluxo de dados e no comportamento geral da aplicação.

4.3. Captura e Processamento de Vídeo

A captura da imagem é feita com a biblioteca OpenCV, utilizando a função `VideoCapture(0)` para acessar a webcam. Cada frame capturado é convertido para o formato RGB, necessário para o funcionamento correto do MediaPipe.

4.4. Detecção dos Landmarks com MediaPipe

O MediaPipe Hands atua como o componente central responsável pela detecção e rastreamento dos pontos anatômicos da mão, viabilizando a extração precisa dos landmarks que servirão de entrada para o classificador. Sua estrutura interna é composta por dois modelos distintos que trabalham de forma complementar. O primeiro é o módulo de Palm Detection, que tem como função identificar rapidamente a presença de uma mão no quadro. Esse modelo é otimizado para localizar a palma com alta eficiência, permitindo que o sistema opere em tempo real mesmo em dispositivos com recursos limitados. Após essa etapa inicial, o Hand Landmark Model entra em ação, sendo responsável por estimar com precisão as posições tridimensionais das 21 articulações da mão, incluindo nós dos dedos, pontas das falanges e segmentos articulados.

O resultado obtido a partir do MediaPipe Hands inclui informações detalhadas sobre o gesto capturado. Para cada uma das 21 articulações detectadas, o sistema retorna as coordenadas normalizadas (x, y, z), que descrevem a posição espacial relativa de cada ponto dentro do quadro. Além disso, são fornecidos dados adicionais, como a indicação de qual mão está sendo rastreada (esquerda ou direita) e a confiança associada ao rastreamento, que expressa o nível de certeza do modelo em relação à detecção realizada. Esses elementos fornecem um panorama completo da configuração manual e permitem avaliar não apenas a posição dos dedos, mas também a estabilidade e a confiabilidade da detecção.

Todas essas informações são essenciais para compor o vetor de entrada utilizado pelo classificador. Cada articulação da mão gera três valores numéricos, correspondentes às coordenadas tridimensionais, resultando em um vetor final composto por 63 elementos. Esse vetor padronizado representa de forma compacta e precisa a estrutura do gesto realizado, servindo como base para o modelo de Random Forest identificar e classificar corretamente a letra da Libras correspondente. A combinação de precisão na detecção, eficiência computacional e simplicidade no formato dos dados torna o MediaPipe Hands uma ferramenta extremamente adequada para aplicações de reconhecimento gestual em tempo real.

4.5. Normalização e Pré-processamento

A normalização dos landmarks é uma etapa essencial para garantir que o modelo de reconhecimento opere de forma consistente, reduzindo variações naturais que poderiam prejudicar

a precisão das classificações. Esse procedimento é particularmente importante porque fatores como a distância da mão em relação à câmera, o tamanho das mãos de diferentes usuários e pequenas movimentações involuntárias durante a captura podem alterar significativamente os valores brutos das coordenadas, gerando discrepâncias entre amostras que representam o mesmo gesto. Sem esse tratamento, o modelo poderia aprender padrões distorcidos ou sensíveis demais às condições externas, resultando em erros de classificação.

Para mitigar esses problemas, foi adotado um procedimento de normalização baseado em princípios geométricos e proporcionais. Primeiramente, todas as coordenadas são ajustadas tendo como referência o ponto-base da mão, geralmente o landmark do punho. Subtrair esse ponto das demais coordenadas desloca a mão para uma posição padronizada no espaço, centralizando a representação e eliminando diferenças decorrentes de posicionamento. Em seguida, os valores são divididos pela distância entre extremidades da mão, como o intervalo entre o pulso e a ponta do dedo médio. Essa etapa reduz o impacto do tamanho físico da mão, tornando os vetores comparáveis entre usuários diferentes. Por fim, os dados passam por uma padronização estrutural, garantindo que todos os vetores tenham o mesmo formato e escala, prontos para serem utilizados pelo classificador.

Esse processo de normalização aumenta significativamente a robustez do modelo, pois torna os vetores de entrada menos dependentes das condições específicas do ambiente e das características físicas dos usuários. Como resultado, o sistema consegue operar em cenários variados (incluindo mudanças de iluminação, distância e ângulo) sem perda perceptível de precisão. Além disso, ao reduzir a variabilidade artificial dos dados, a normalização contribui para que o modelo aprenda de forma mais eficiente os padrões realmente relevantes, aprimorando sua capacidade de generalização e tornando o reconhecimento das letras mais confiável em aplicações reais.

4.6. Classificação com Random Forest

O modelo utilizado para a etapa de classificação foi o `RandomForestClassifier`, configurado com 200 árvores de decisão. A escolha dessa abordagem se deu por uma combinação de fatores que a tornam especialmente adequada para o problema em questão. Entre suas características mais relevantes destacam-se o bom desempenho em datasets pequenos, a robustez diante de ruídos e imperfeições nos dados e sua alta capacidade de generalização, o que permite identificar corretamente padrões mesmo em situações ligeiramente diferentes das observadas durante o treinamento.

O Random Forest é particularmente eficiente em projetos como este porque trabalha com múltiplas árvores treinadas sobre subconjuntos distintos dos dados e dos atributos, o que reduz a sensibilidade a outliers e diminui substancialmente o risco de overfitting. Isso significa que, mesmo quando o dataset não é extremamente amplo (como ocorre em bases geradas manualmente, a partir de landmarks), o modelo consegue abstrair os padrões relevantes das posições anatômicas da mão sem se prender a variações acidentais, iluminando o equilíbrio entre simplicidade e precisão.

Após o treinamento inicial, que é realizado de forma offline, o modelo é armazenado e carregado durante a execução do sistema. Esse procedimento garante que a etapa de reconhecimento ocorra de maneira leve e eficiente, já que não é necessário treinar o classificador novamente a cada uso. A predição das letras é feita em poucos milissegundos, mesmo em hardware simples, permitindo que o sistema funcione plenamente em tempo real. Essa velocidade de inferência é um dos elementos centrais para a boa experiência do usuário, pois possibilita que gestos sejam interpretados instantaneamente, sem atrasos perceptíveis, reforçando o caráter

dinâmico e interativo da aplicação.

4.7. Interface Web

A interface web foi desenvolvida para tornar a aplicação acessível, intuitiva e visualmente agradável. Ela é composta por três telas, que são: Tela Inicial, Modo Escrever e Modo Quiz.

4.7.1. Tela Inicial

A tela inicial da aplicação atua como ponto de entrada para o sistema, oferecendo ao usuário uma visão clara e organizada das funcionalidades disponíveis. Nessa interface, o usuário pode escolher entre o Modo Escrever e o Modo Quiz, permitindo acesso rápido e direto às principais ferramentas do sistema. A disposição centralizada dos elementos e o uso de botões de navegação bem definidos contribuem para uma experiência intuitiva, facilitando o entendimento imediato do funcionamento da aplicação, mesmo para indivíduos com pouca familiaridade com tecnologias digitais.

Além disso, o design minimalista e o contraste visual buscam garantir legibilidade e acessibilidade, características essenciais em contextos educacionais e assistivos. A tela inicial também desempenha papel importante na orientação do usuário, funcionando como um ambiente de transição entre o acesso ao sistema e o início das atividades de reconhecimento, contribuindo para uma interação mais fluida e organizada.

4.7.2. Modo Escrever

O Modo Escrever foi desenvolvido com o objetivo de traduzir gestos em Libras para texto de maneira simples, fluida e intuitiva, permitindo que o usuário pratique e aprenda as letras do alfabeto de forma dinâmica. Nesse modo, o usuário posiciona a mão diante da webcam, e o sistema realiza o reconhecimento da letra correspondente com base nos landmarks capturados em tempo real. Assim que o gesto é identificado, a letra reconhecida é automaticamente inserida em uma caixa de texto na interface do sistema, onde o usuário pode editar, apagar, selecionar ou combinar os caracteres para formar palavras e frases completas. Esse recurso transforma o uso da Libras em um processo interativo de escrita, possibilitando tanto o treinamento das letras quanto a criação de pequenos trechos de texto exclusivamente a partir dos sinais manuais.

Durante o desenvolvimento desse modo, alguns desafios importantes precisaram ser solucionados. Um dos principais consistiu em garantir a fluidez da escrita, evitando que variações rápidas ou instabilidades na posição da mão resultassem em múltiplos caracteres indesejados. Além disso, foi necessário implementar mecanismos de suavização ou atraso para impedir que mudanças bruscas ou interpretações instantâneas equivocadas afetassem a experiência do usuário. Para lidar com essas situações, foi desenvolvida uma lógica de “detecção estável”, na qual a letra só é registrada após permanecer consistentemente reconhecida por um pequeno intervalo de tempo. Esse procedimento permitiu reduzir ruídos, aumentar a precisão da escrita e tornar o processo mais natural, aproximando-se do ritmo de digitação manual, porém utilizando gestos em Libras.

4.7.3. Modo Quiz

O modo Quiz foi desenvolvido com foco educacional, visando proporcionar ao usuário uma forma dinâmica, interativa e prática de aprender os sinais do alfabeto em Libras. Nesse modo,

o sistema seleciona aleatoriamente uma letra do conjunto de sinais estáticos e apresenta o desafio ao usuário, que deve tentar reproduzir o gesto correspondente em frente à webcam. Em seguida, o sistema analisa a posição da mão e compara os landmarks capturados com o padrão da letra sorteada. Caso o sinal realizado corresponda corretamente ao gesto esperado, um feedback visual em verde é exibido na tela, indicando acerto; caso contrário, o retorno aparece em vermelho, sinalizando que o gesto não foi reconhecido como correspondente. Após cada tentativa, o quiz avança automaticamente para a próxima letra, permitindo que o usuário pratique sucessivamente diferentes sinais de forma contínua.

Essa dinâmica de desafio e resposta contribui diretamente para o processo de aprendizagem ativa, estimulando a memorização e reforçando a prática repetitiva dos sinais. Além disso, o modo Quiz é especialmente útil em ambientes educacionais, pois transforma o estudo do alfabeto em uma atividade lúdica e motivadora, facilitando o engajamento de estudantes iniciantes na Libras e promovendo o desenvolvimento gradual da habilidade de reconhecer e executar corretamente as configurações manuais.

4.7.4. Imagens da Aplicação

A interface inicial do sistema é apresentada na Figura 2, onde se observa que o layout é composto por uma estrutura simples e intuitiva. Nessa tela, encontram-se os botões de navegação para os modos disponíveis, permitindo acesso direto às funcionalidades de reconhecimento. O design prioriza a clareza visual e a facilidade de uso, especialmente para usuários iniciantes, reforçando o caráter inclusivo da aplicação.

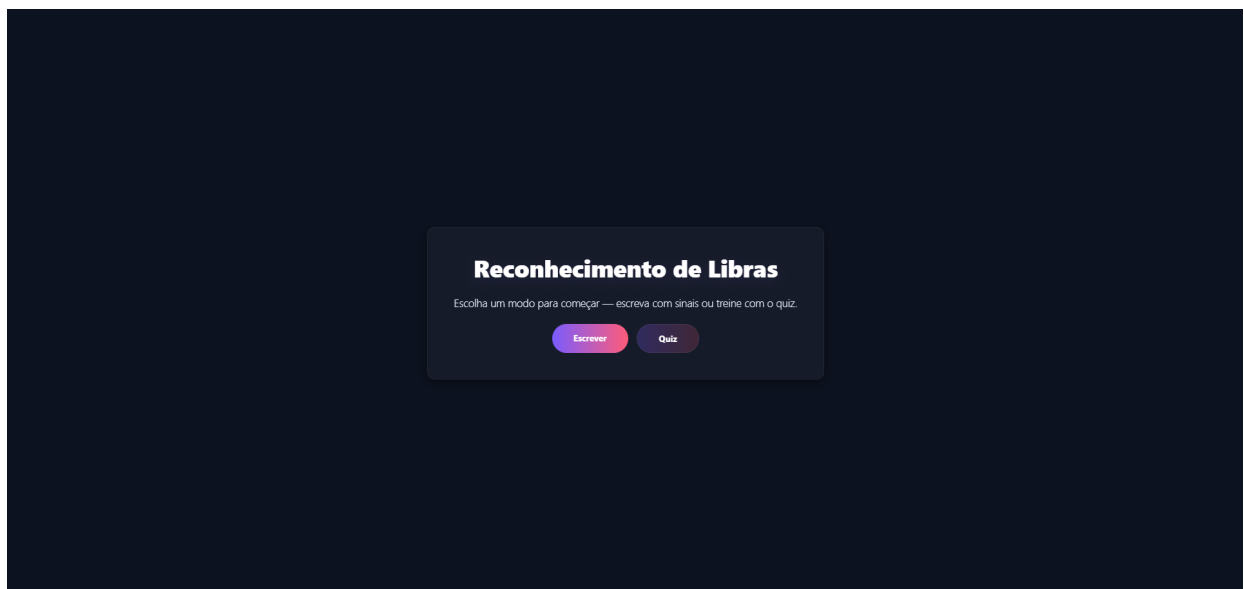


Figura 2. Tela Inicial

A interface do Modo Escrever, ilustrada na Figura 3, demonstra o ambiente no qual o sistema converte gestos reconhecidos em texto. Nessa tela, observa-se a área destinada à visualização da captura da webcam, bem como a caixa de texto onde os caracteres identificados são inseridos automaticamente. Além disso, são exibidos botões funcionais que permitem ao usuário editar o texto produzido, incluindo ações como apagar caracteres, inserir espaços e limpar

o conteúdo. Essa organização visual favorece o uso contínuo e a prática da escrita em Libras de forma dinâmica e interativa.

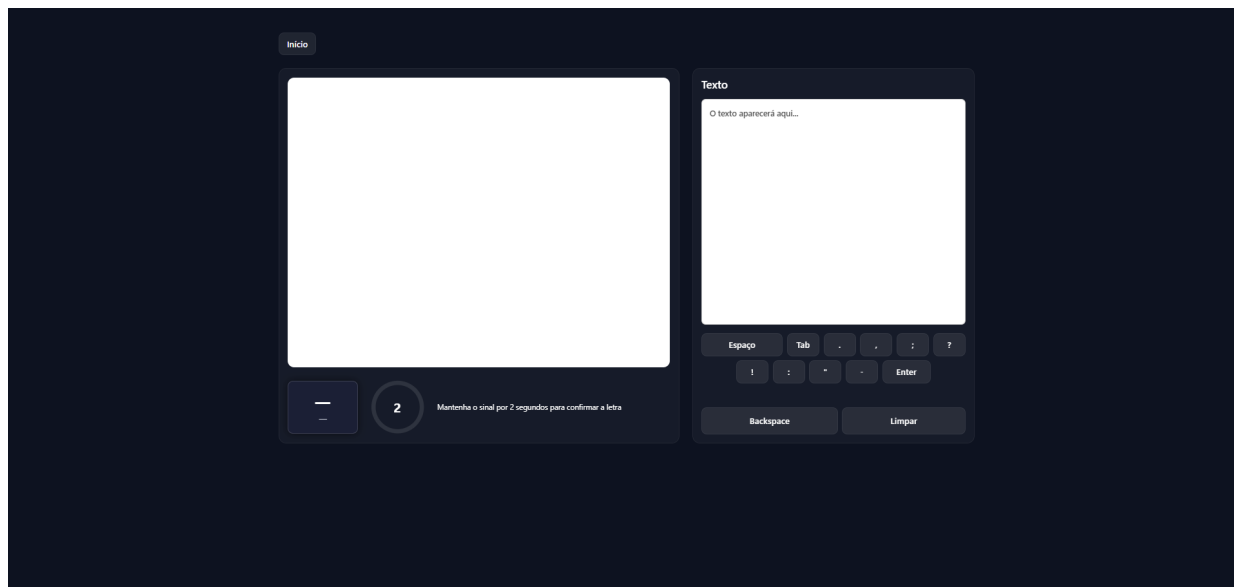


Figura 3. Tela do Modo Escrever

A Figura 4 apresenta a interface do Modo Quiz, na qual o usuário é desafiado a reproduzir corretamente sinais do alfabeto em Libras. É possível observar o destaque dado à letra alvo e ao indicador visual de acerto ou erro, elementos essenciais para o feedback imediato durante a execução do exercício. A disposição dos componentes visa facilitar a interação e manter o usuário focado na atividade, contribuindo para um processo de aprendizagem mais engajador e eficiente. Essa configuração reforça o caráter educacional da ferramenta, permitindo prática contínua e avaliação instantânea do desempenho.

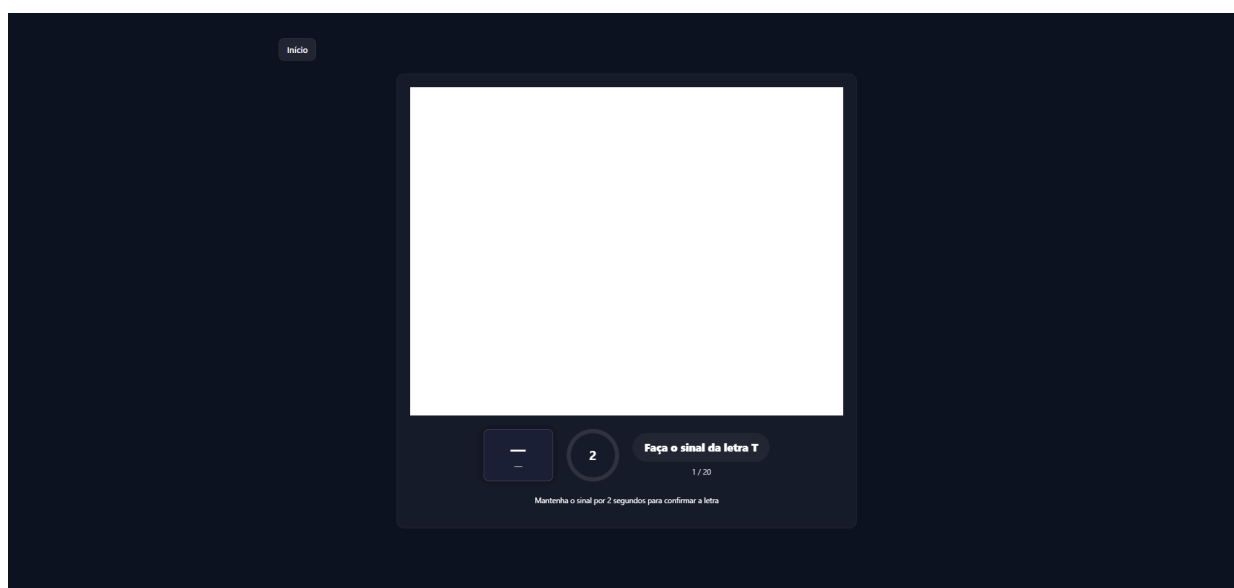


Figura 4. Tela do Modo Quiz

5. Resultados e Discussão

Os testes realizados demonstraram um desempenho consistente e satisfatório, tanto no ambiente controlado de avaliação quanto no uso prático em tempo real. O modelo Random Forest apresentou acurácia elevada no conjunto de teste, indicando boa capacidade de generalização e manutenção de desempenho estável durante a execução contínua com a webcam. Em situações reais, o sistema se mostrou eficiente em reconhecer diferentes configurações manuais, mesmo diante de variações de posição, distância e pequenos ajustes de orientação da mão, reforçando a adequação do classificador escolhido ao tipo de dado utilizado.

A Tabela 2 apresenta um resumo das principais métricas obtidas durante os testes, considerando acurácia geral e taxa média de acertos por classe, evidenciando o desempenho satisfatório do sistema para uso em tempo real.

Tabela 2. Resumo dos resultados obtidos nos testes

Métrica	Resultado Obtido
Acurácia geral	92,40%
Classes com maior taxa de erro	M, N, R
Sensibilidade a iluminação	Baixa
Execução em hardware sem GPU dedicada	Estável

O funcionamento em tempo real foi observado como um dos pontos mais positivos da solução, permitindo que as letras fossem reconhecidas e exibidas instantaneamente na interface. Mesmo em computadores com hardware modesto, sem placas gráficas dedicadas, a inferência permaneceu fluida, resultado direto da leveza computacional proporcionada pelo uso de landmarks tridimensionais e do baixo custo de processamento do Random Forest. Adicionalmente, o sistema apresentou baixa sensibilidade a variações de iluminação e a diferentes tipos de fundo, demonstrando robustez importante para seu uso em ambientes cotidianos, como salas de aula, residências ou laboratórios universitários.

Outro aspecto relevante refere-se à simplicidade do classificador, que favorece não apenas o desempenho, mas também a interpretabilidade. Ao contrário de modelos mais complexos, como redes neurais profundas, o Random Forest possibilita analisar a importância das características e o comportamento das árvores individuais, facilitando ajustes futuros, depuração de erros e compreensão dos padrões aprendidos. Essa característica torna o sistema mais acessível pedagogicamente, permitindo que estudantes e pesquisadores compreendam seu funcionamento mesmo sem conhecimento avançado em deep learning.

A integração das funcionalidades pedagógicas, como o Modo Escrita e o Modo Quiz, também se destacou como diferencial da proposta. Além de reconhecer gestos, o sistema oferece ferramentas práticas que estimulam o aprendizado ativo da Libras, tornando-se não apenas um recurso técnico, mas também didático. O Modo Escrita permite que o usuário produza palavras e frases utilizando apenas gestos, enquanto o Modo Quiz promove uma experiência de treino interativa, reforçando a memorização das letras. Esses modos complementam o reconhecimento automático e ampliam as possibilidades de aplicação em contextos educacionais, reforçando o caráter inclusivo e formativo da ferramenta.

Ao analisar os resultados em conjunto, observa-se que a abordagem baseada em landmarks associada ao classificador Random Forest mostrou-se adequada para o reconhecimento das letras estáticas da Libras. A escolha metodológica permitiu conciliar simplicidade computacional,

eficiência e boa acurácia sem exigir infraestruturas avançadas. Entretanto, também foram identificadas limitações: o sistema é restrito a sinais estáticos e não contempla gestos que envolvem movimento contínuo, exigindo metodologias mais complexas para expansão futura. Além disso, embora os landmarks reduzam a sensibilidade à iluminação, situações extremas de sombra ou recortes incorretos da mão podem impactar o desempenho em certos cenários.

Mesmo assim, considerando os objetivos do projeto, os resultados obtidos confirmam a viabilidade da solução proposta, demonstrando que é possível desenvolver um sistema funcional, leve e acessível capaz de atuar como ferramenta de apoio tanto para comunicação quanto para ensino de Libras. O bom desempenho em condições reais indica que o protótipo pode servir como base para estudos mais avançados, como reconhecimento de gestos dinâmicos, construção de frases completas e expansão para outros elementos linguísticos da Libras.

A Figura 5 apresenta o funcionamento do Modo Escrever durante o reconhecimento da letra “A” em tempo real. Nela, é possível observar a captura da mão pela webcam com os landmarks detectados pelo MediaPipe, representados por pontos e conexões coloridas que indicam a posição dos dedos. A interface exibe também o nível de confiança do classificador, que alcança 98,5%, e o temporizador circular responsável por confirmar o gesto após um curto período de estabilidade. À direita, a letra reconhecida é inserida automaticamente na área de texto, evidenciando a integração entre o reconhecimento gestual e a escrita na aplicação. Essa configuração demonstra a natureza interativa e imediata do sistema.

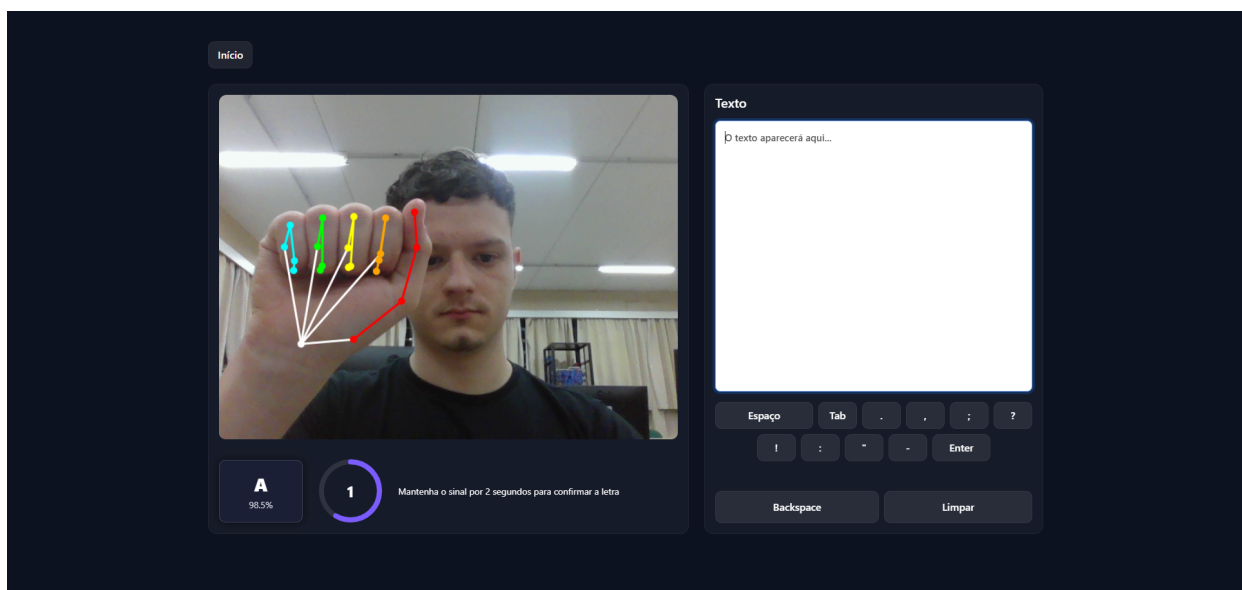


Figura 5. Modo Escrever identificando a letra A

A Figura 6 apresenta o funcionamento do Modo Quiz, no qual o usuário deve reproduzir corretamente o sinal da letra indicada pelo sistema (D). Na imagem, observa-se a captura da mão pela webcam com os landmarks detectados pelo MediaPipe, permitindo visualizar o posicionamento dos dedos durante a execução do gesto. A interface exibe o nível de confiança do classificador, que neste caso é de 82,5%, além do temporizador circular e da instrução “Faça o sinal da letra D”, indicando que o gesto realizado ainda não corresponde ao sinal esperado. Essa configuração evidencia o papel do feedback imediato no processo de aprendizagem, permitindo ao usuário ajustar o movimento até atingir a configuração correta.

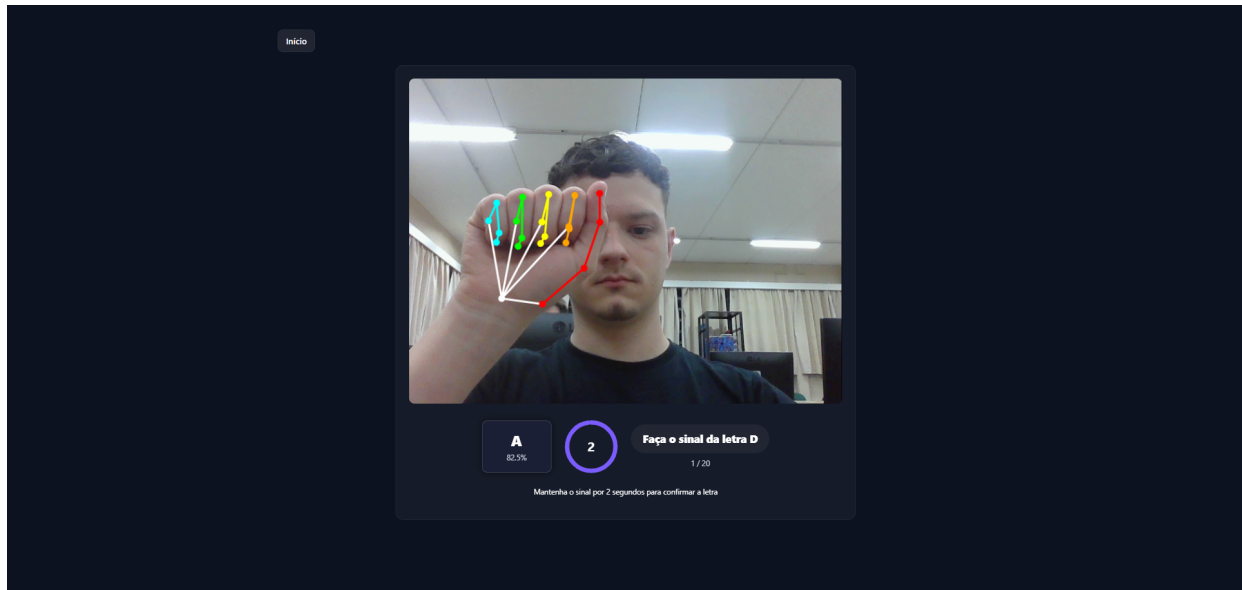


Figura 6. Modo Quiz não identificando a letra D

6. Conclusão e Trabalhos Futuros

O presente trabalho desenvolveu um sistema capaz de reconhecer, em tempo real, as letras estáticas do alfabeto da Língua Brasileira de Sinais (Libras) utilizando visão computacional e aprendizado de máquina. A proposta buscou contribuir para a acessibilidade e a inclusão, oferecendo uma ferramenta tecnológica que apoia a comunicação entre surdos e ouvintes.

O sistema foi construído a partir de um dataset próprio e implementado com o uso do MediaPipe Hands para detecção dos landmarks da mão, combinado ao modelo Random Forest para classificação. Essa abordagem apresentou desempenho satisfatório, alcançando acurácia superior a 90% e funcionamento estável em diferentes condições de uso, sem necessidade de equipamentos específicos. A interface web, com os modos Escrita e Quiz, demonstrou potencial tanto como recurso assistivo quanto como ferramenta educacional e de aprendizagem.

Apesar dos resultados positivos, o sistema ainda apresenta limitações, sobretudo quanto ao reconhecimento de letras que envolvem movimento dinâmico, como H, J, K, X, Y e Z. Mesmo assim, o protótipo confirmou sua viabilidade técnica e social, destacando-se pela acessibilidade, baixo custo e facilidade de uso.

Além disso, este trabalho abre espaço para importantes evoluções futuras. Entre elas, destaca-se a inclusão das letras dinâmicas da Libras por meio de modelos capazes de analisar sequências temporais, como LSTM ou GRU; a ampliação do sistema para o reconhecimento de palavras e frases completas, permitindo aplicações em comunicação assistiva e tradução automática; e o desenvolvimento de um aplicativo móvel utilizando tecnologias como TensorFlow Lite ou MediaPipe Mobile, tornando a solução mais portátil e acessível.

A continuidade dessas pesquisas poderá resultar em sistemas mais completos e robustos, ampliando o impacto social da ferramenta e fortalecendo o uso da inteligência artificial como suporte à comunidade surda no Brasil.

7. Referências Bibliográficas

[Bishop 2006] [BRASIL 2002] [BRASIL 2005] [BRASIL 2015] [Breiman 2001]
[Furtado e de Oliveira 2021] [Goodfellow et al. 2016] [de Educação de Surdos (INES) 2023]

[Neto 2019] [de Quadros e Karnopp 2004] [de Quadros 2009] [Silva 2021] [Szeliski 2022]
[World Health Organization e World Bank 2011] [Krizhevsky et al. 2012] [LeCun et al. 2015]
[Lugaresi et al. 2019]

Referências

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.
- BRASIL (2002). Lei nº 10.436, de 24 de abril de 2002. dispõe sobre a língua brasileira de sinais – libras. Diário Oficial da União.
- BRASIL (2005). Decreto nº 5.626, de 22 de dezembro de 2005. regulamenta a lei nº 10.436/2002. Diário Oficial da União.
- BRASIL (2015). Lei nº 13.146, de 6 de julho de 2015. estatuto da pessoa com deficiência. Diário Oficial da União.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.
- de Educação de Surdos (INES), I. N. (2023). *Educação de Surdos em Perspectiva Bilíngue*. Rio de Janeiro. E-book.
- de Quadros, R. M. (2009). *Língua Brasileira de Sinais I*. Editora UFSC, Florianópolis. E-book.
- de Quadros, R. M. e Karnopp, L. B. (2004). *Língua de Sinais Brasileira: Estudos Linguísticos*. Artmed, Porto Alegre.
- Furtado, S. L. e de Oliveira, J. (2021). Computer vision and neural networks for libras recognition. In *Anais do Workshop de Visão Computacional (WVC)*, number 17, São Paulo. SBC.
- Goodfellow, I., Bengio, Y., e Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge.
- Krizhevsky, A., Sutskever, I., e Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.
- LeCun, Y., Bengio, Y., e Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.
- Lugaresi, C. et al. (2019). Mediapipe: A framework for building perception pipelines. Versão alternativa em HTML: <https://ar5iv.labs.arxiv.org/html/1906.08172>.
- Neto, V. C. L. (2019). Reconhecimento de sinais do alfabeto da libras utilizando visão computacional.
- Silva, B. V. L. (2021). Reconhecimento de sinais da libras por visão computacional.
- Szeliski, R. (2022). *Computer Vision: Algorithms and Applications*. Springer, Cham, 2 edition. Versão eletrônica (draft).
- World Health Organization e World Bank (2011). World report on disability. Versão em português: Relatório mundial sobre a deficiência. São Paulo: SEDPcD, 2012.