Sistema Inteligente na Interpretação da Língua Brasileira de Sinais Em Tempo Real

Elisa Ayumi Masasi de Oliveira
Instituto de Informática
Universidade Federal de Goiás
Goiânia, Goiás
ayumi@discente.ufg.br

Evellyn Nicole Machado Rosa
Instituto de Informática
Universidade Federal de Goiás
Goiânia, Goiás
nicole@discente.ufg.br

Iago Alves Brito
Instituto de Informática
Universidade Federal de Goiás
Goiânia, Goiás
iagoalves@discente.ufg.br

Samuel França da Costa Pedrosa
Escola de Engenharia Elétrica, Mecânica e de Computação
Universidade Federal de Goiás
Goiânia, Goiás
samuelpedrosa@discente.ufg.br

***Abstract*—A comunicação é o meio fundamental pelo qual ocorre o desenvolvimento social dos indivíduos, sucedendo-se através dela a interação e integração dos seres humanos entre si, estabelecendo pilares para a manutenção da sociedade e a internalização de padrões sociais (BERGER, 1973). A possibilidade de se comunicar, portanto, é o que permite aos indivíduos trocarem informações e estabelecerem relações, necessitando entender e ser entendido para que ocorra esse processo. Dessa forma, para que indivíduos com diferentes tipos de deficiência auditiva e surdez se comuniquem, a utilização da língua de sinais assume papel fundamental, sendo ela o motor para o processo de socialização destes indivíduos. Como forma de auxiliar a comunicação entre falantes e não-falantes da Língua Brasileira de Sinais (LIBRAS), o presente trabalho apresenta um sistema capaz de realizar a interpretação em tempo real de Libras através do uso de técnicas baseadas em Inteligência Artificial, um importante passo para a inserção plena da população brasileira cuja principal forma de comunicação é a linguagem de sinais.**

***Index Terms*—Libras, Machine Learning, Visão Computacional, Redes Neurais Profundas.**

I. INTRODUÇÃO

Libras é a sigla para Língua Brasileira de Sinais, uma língua de modalidade gestual-visual em que é possível se comunicar por meio de gestos, expressões faciais e corporais. Desde 24 de abril de 2002, Libras se tornou uma língua oficial brasileira.

Essa linguagem é um meio de comunicação muito importante para a comunidade não só de surdos e mudos, mas também para qualquer pessoa que irá interagir com esses indivíduos. Exemplo disso é o aumento no aparecimento de intérpretes em vídeos e transmissões ao vivo. No entanto, essa interpretação nem sempre é de fácil acesso para a população.

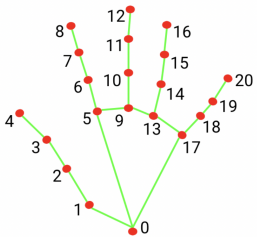
Diante desse cenário, a tecnologia de visão computacional emerge como uma solução promissora para superar esses desafios de comunicação. Este artigo propõe-se a explorar um projeto inovador que utiliza a visão computacional para identificar e traduzir sinais de Libras em tempo real. Ao aproveitar os avanços na detecção de padrões gestuais e na interpretação de movimentos das mãos, esse projeto busca tornar a comunicação em Libras acessível e compreensível para um público mais amplo

II. FUNDAMENTOS

Para que seja possível realizar a tradução, é necessário utilizar mecanismos de detecção de rosto, rastreamento de mãos e reconhecimento de Hand Landmarks*. Como por exemplo o reconhecimento e rastreamento do framework MediaPipe que utiliza pipelines de machine learning. Ou seja, uma ordem de processos protocolados utilizando aprendizado de máquina. No caso foram utilizados, pelo MediaPipe, mais de trinta mil imagens de mãos com características e fundos de imagem diferentes para treinamento da detecção da mão.

Em um vídeo, utiliza-se um intervalo de frames como imagens que passam por processos de visão computacional como rotação, redimensionamento, normalização e conversão de cores para melhorar a eficiência da detecção da mão na imagem. Em seguida, é detectado se é uma mão esquerda ou direita e seus 21 pontos-chaves.

Os 21 pontos-chaves são mostrados na Figura 1:



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

Figura 1: Hand Landmarks.

Cada ponto-chave possui três coordenadas, x, y e z, sendo o "x" e "y" coordenadas relacionadas com a altura e largura baseadas em pixels da imagem, e o "z" como profundidade variando de 0.0 a 1.0 com referência no punho (ponto-chave zero).

O processo também cria uma bounding box (caixa delimitadora) da mão para o rastreamento no frame seguinte, é

utilizado como base o bounding box do frame anterior para o rastreamento e localização no frame atual, evitando chamadas consecutivas da detecção de mão na imagem inteira, o que tornaria o processo mais custoso computacionalmente. Convocando apenas a chamada do rastreamento de mão novamente quando o rastreamento for interrompido.

Além das ferramentas do MediaPipe, os recursos de Visão Computacional contarão com a arquitetura robusta de uma Rede Neural Recorrente (LSTM) para rotulação dos sinais que são reconhecidos. Dessa forma o algoritmo completo conseguirá, a partir de um vídeo ou streaming, detectar e rastrear as mãos e articuladores (ombros, braços e rostos) e verificar qual o sinal rotulado mais compatível com o que está sendo representado.

III. METODOLOGIA

A utilização das soluções do framework MediaPipe foram de grande relevância para a concretização deste estudo. Posto que suas ferramentas foram necessárias para todo o processo de desenvolvimento que antecede o treinamento. Desde a criação do script que realiza o tracking facial, de pose, dos braços esquerdo e direito, e então das landmarks da mão, até a consolidação do dataset.

Dataset esse, que pôde ser considerado o maior desafio enfrentado na execução deste estudo. Sendo considerado um dos fatores mais importantes no treinamento de um modelo com resultados satisfatórios, o dataset representou um grande obstáculo no desenvolvimento do presente estudo, devido à falta de conjuntos de dados em português. Há um vasto cenário de dados quando se trata do reconhecimento do alfabeto, entretanto, com o objetivo de que ao final desse processo seja concebida uma ferramenta que auxilie a comunicação e interpretação, o dataset necessário precisaria conter palavras e frases, o que não foi encontrado de maneira categórica. Levando assim, o grupo a criar um conjunto de dados limitado, mas que atendesse às expectativas do estudo.

A utilização das soluções do framework MediaPipe proporcionou uma base sólida para a aplicação de técnicas de aprendizado de máquina, como a construção de Redes Neurais Recorrentes (RNNs), um componente essencial para a fase de treinamento do modelo. As RNNs, especialmente aquelas com variações como Long Short-Term Memory (LSTM) ou Gated Recurrent Unit (GRU), são particularmente vantajosas para lidar com dados sequenciais, como os gerados pelas landmarks das mãos capturadas pelo MediaPipe. Essa abordagem permitiu a consideração das dependências temporais e a relação entre os diferentes pontos em sequência, fatores cruciais para a interpretação eficaz dos gestos. Portanto, a integração da detecção e rastreamento de landmarks do MediaPipe com as técnicas de processamento de sequência, como as RNNs, desempenhou um papel central na criação de um modelo capaz de interpretar os gestos com maior precisão e contexto.

Por outro lado, a estratégia de integração das Redes Neurais Recorrentes (RNNs) também foi um componente crucial para a construção do modelo. Ao aplicar filtros sobre as imagens das mãos, as RNNs foram capazes de extrair informações vitais, como bordas e texturas, permitindo que o modelo identificasse padrões complexos presentes nos dados de entrada. A interação entre as RNNs e as ferramentas do MediaPipe possibilitou um nível significativo de precisão e contexto na interpretação dos gestos das mãos.

IV. RESULTADOS

Para a avaliação na construção do modelo, foi utilizada a acurácia, que alcançou um desempenho de 1,0. Esse alto valor, no entanto, se deve ao fato de haverem poucas amostras disponíveis para teste, o que somado ao fato de haver um único articulador, se traduziu numa base de dados limitada ao contexto de pesquisa. A matriz de confusão gerada abaixo expressa o resultado obtido.

		Predito	
		Positivo	Negativo
Real	Positivo	1	0
	Negativo	0	2

Tabela 1: Matriz de confusão para a classe "Oi"

		Predito	
		Positivo	Negativo
Real	Positivo	2	0
	Negativo	0	1

Tabela 2: Matriz de confusão para a classe "Eu Te Amo"

Além da matriz de confusão como métrica, o bom desempenho na inferência em tempo real [4] serviu como pilar para determinar os objetivos deste estudo como concluídos, haja vista que nas fases de treinamento propostas o sistema conseguiu classificar corretamente os sinais.

V. CONCLUSÃO E PRÓXIMOS PASSOS

Ao concluir o presente estudo, foi possível perceber que apesar das limitações encontradas, foi possível gerar um produto funcional, que age em prol da acessibilidade para a inserção plena da população brasileira cuja principal forma de comunicação é a linguagem de sinais

Para próximos estudos, recomenda-se o desenvolvimento de uma base de dados mais robusta para o treinamento e validação do modelo, bem como o teste com uma grande quantidade de falantes de Libras de forma a verificar a real performance do modelo. Posteriormente, recomenda-se a instauração da solução em locais públicos para possibilitar a comunicação entre falantes e não falantes de Libras, sempre visando a inclusão e acessibilidade desta camada da população do Brasil.

REFERENCES

[1] R Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91.

[2] RODRIGUES, Ailton José. V-LIBRASIL: uma base de dados com sinais na língua brasileira de sinais (Libras). 2021. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Pernambuco, Recife, 2021.

[3] GOOGLE DEVELOPERS. Mediapipe: Hand Landmarker. 2023. Disponível em: ¡https://developers.google.com/mediapipe/solutions/vision/hand¿$andmarker >$ $.Acessoem$: $16ago.2023$.

[4] [MASASI, Elisa]. [Trabalho Final]. 16 de agosto, 2023. Disponível em: ¡https://drive.google.com/drive/folders/1e1pyXADQDn$_V Fkm5BI0F7Wml78rMpS8g?usp = sharing > .Acessoem$: 16ago.2023.

[5] NOCHNACK, Nick. Action Detection for Sign Language. Disponível em: ¡https://github.com/nicknochnack/ActionDetectionforSignLanguage¿. Acesso em: 16 agosto, 2023.