Desenvolvimento de um Identificador de Sinais de LIBRAS usando Visão Computacional

Development of a LIBRAS Sign Identifier using Computer Vision

Luana Cristina Guerreiro Campos[1], Gustavo Henrique Paetzold[2]

RESUMO

A Libras, Língua Brasileira de Sinais, é essencial para a comunicação da comunidade surda no Brasil. Sua rica diversidade de gestos e expressões apresenta desafios para o reconhecimento automático por meio de sistemas computacionais. O principal objetivo deste trabalho foi criar um conjunto de dados e utilizá-los em um modelo de aprendizado de máquina capaz de reconhecer os gestos em Libras. Para concretizar esse objetivo, vídeos do YouTube foram utilizados para formar um conjunto com 22.000 imagens anotadas. A biblioteca MediaPipe, conhecida por suas capacidades avançadas de visão computacional, foi empregada para detectar as articulações das mãos nas imagens. Utilizando esse conjunto de dados, um modelo de aprendizado de máquina foi treinado e alcançou um F-score médio de 89% na identificação dos gestos em Libras. Em conclusão, o trabalho demonstra o potencial do aprendizado de máquina em promover a acessibilidade em Libras, com perspectivas futuras incluindo a capacidade de reconhecer sinais mais complexos.

PALAVRAS-CHAVE: Aprendizado de Máquina; Libras; Visão Computacional.

ABSTRACT

Libras, the Brazilian Sign Language, is essential for the communication of the deaf community in Brazil. Its rich diversity of gestures and expressions presents challenges for automatic recognition through computational systems. The main objective of this work was to create a dataset and use it in a machine learning model capable of recognizing Libras gestures. To fulfill this aim, YouTube videos were sourced to compile a dataset containing 22,000 annotated images. The MediaPipe library, renowned for its advanced computer vision capabilities, was utilized to detect hand articulations in the images. Employing this dataset, a machine learning model was trained, achieving an average F-score of 89% in identifying Libras gestures. In conclusion, this work demonstrates the potential of machine learning in promoting accessibility in Libras, with future prospects encompassing the ability to recognize more complex signs.

KEYWORDS: Machine Learning; Libras; Computer Vision.

[1] Voluntária. Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil. E-mail: luanacampos@alunos.utfpr.edu.br. ID Lattes: 4780074821595493.

[2] Docente no curso de Engenharia de Computação. Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil. E-mail: ghpaetzold@utfpr.edu.br. ID Lattes: 3576463426605379.

INTRODUÇÃO

A Língua Brasileira de Sinais (Libras), é uma língua de modalidade gestual-visual onde é possível se comunicar através de gestos, expressões faciais e corporais. O alfabeto manual é uma componente da Libras e de muitas outras línguas de sinais pelo mundo. Consiste em uma série de gestos manuais que correspondem às letras do alfabeto escrito, frequentemente utilizado para soletrar nomes próprios ou termos técnicos (Cristiano, 2017).

O aprendizado de máquina está se mostrando essencial para melhorar a comunicação e a acessibilidade com Libras. Com a utilização de algoritmos avançados, sistemas podem ser desenvolvidos para reconhecer sinais com acurácia, facilitando a integração da Libras em plataformas digitais. Nesse contexto, entra a visão computacional, que é uma subárea da inteligência artificial focada em capacitar máquinas para interpretar e entender o conteúdo visual, como imagens e vídeos.

Em outros estudos, a tarefa de reconhecer sinais foi realizada utilizando o Kinect, um dispositivo de entrada de detecção de movimento desenvolvido pela Microsoft. O Kinect capta dados de profundidade e pode rastrear movimentos do corpo, tornando-se uma escolha popular para projetos relacionados ao reconhecimento de gestos (LANG, 2012). A decisão de utilizar a visão computacional decorreu de sua ampla aplicabilidade. Ao contrário de soluções baseadas em hardware específico, como o Kinect, a visão computacional pode ser implementada usando câmeras comuns, tornando a solução mais acessível.

O objetivo principal deste trabalho foi desenvolver um modelo de aprendizado de máquina que, ao receber a imagem do sinalizador[3] como entrada, é capaz de extrair as articulações da mão relevantes e, em seguida, classificar a letra do alfabeto correspondente. Para alcançar esse objetivo, ferramentas como Python, MediaPipe e Scikit-learn foram fundamentais no desenvolvimento do modelo.

METODOLOGIA

COLETA DOS DADOS

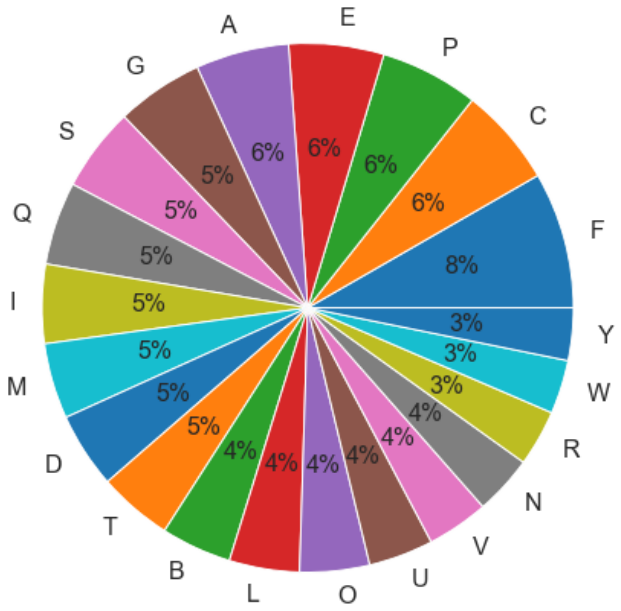
Um dos grandes desafios enfrentados na pesquisa e desenvolvimento de tecnologias voltadas para línguas de sinais é a escassez de dados anotados disponíveis. Como solução para este impasse, recorreu-se à obtenção de dados por meio de vídeos disponíveis no YouTube, uma abundante fonte de conteúdo em Libras, gerado tanto por educadores quanto por membros da comunidade surda.

Após a coleta, o passo seguinte foi a anotação desses dados, para cada um dos frames dos vídeos, a área de destaque de cada sinal foi demarcada. Para essa tarefa, utilizou-se a ferramenta CVAT (Computer Vision Annotation Tool).

Durante o processo de construção do conjunto de dados, 22 mil frames foram anotados. O conjunto de dados possui 17 sinalizadores distintos, assegurando uma diversidade significativa para análises posteriores. A distribuição detalhada por classe pode ser visualizada na Figura 1.

[3] Pessoa que realiza o sinal.

Figura 1 - Distribuição do conjunto de dados por classe



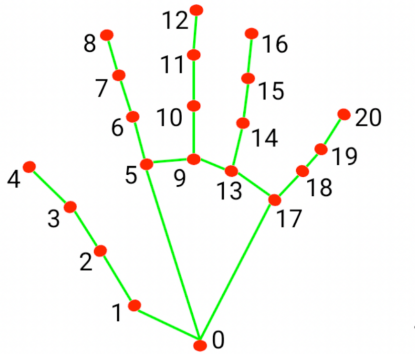
Fonte: Autoria própria, 2023.

PROCESSAMENTO DOS DADOS

MediaPipe é uma biblioteca de código aberto desenvolvida pelo Google, focada em facilitar o desenvolvimento de aplicações em visão computacional, oferecendo ferramentas para detecção de mãos, rastreamento de poses, reconhecimento facial, entre outras funcionalidades. É otimizada para desempenho em tempo real, tornando-se uma solução popular para projetos que requerem processamento de imagens e vídeos com baixa latência (Lugaresi, 2019).

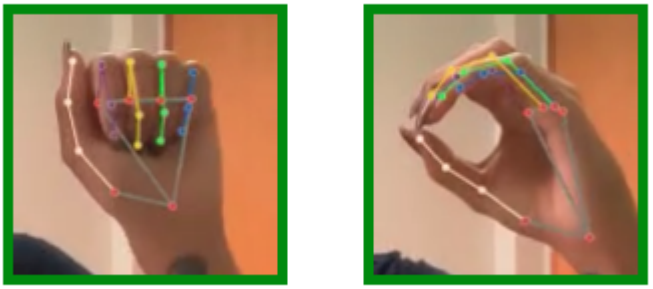
Com o auxílio da biblioteca MediaPipe, foi possível extrair pontos-chaves das mãos. Cada um desses pontos, representa uma parte distinta da mão, seja uma articulação ou a ponta de um dedo, conforme ilustrado nas Figuras 2 e 3.

Figura 2 - Pontos-chaves da biblioteca MediaPipe



Fonte: Lugaresi, 2019.

Figura 3 - Exemplos de sinais



Fonte: Autoria própria, 2023.

TREINAMENTO DO MODELO DE APRENDIZADO DE MÁQUINA

O processo de treinamento do modelo envolveu a utilização da técnica de Máquinas de Vetores de Suporte. O SVM é um método de aprendizado supervisionado em *machine learning* que é utilizado para classificação e regressão. A principal ideia do SVM é encontrar um hiperplano que melhor separa as classes de dados em um espaço multidimensional (Cortes, 1995).

A entrada para o modelo é uma representação vetorial das articulações da mão capturadas a partir da imagem do sinalizador. Cada articulação é identificada por suas coordenadas (x, y, z) no espaço, resultando em um vetor multidimensional que captura a configuração espacial da mão no momento do sinal. A saída do modelo, é a classificação do sinal correspondente, representando uma letra do alfabeto manual da Libras.

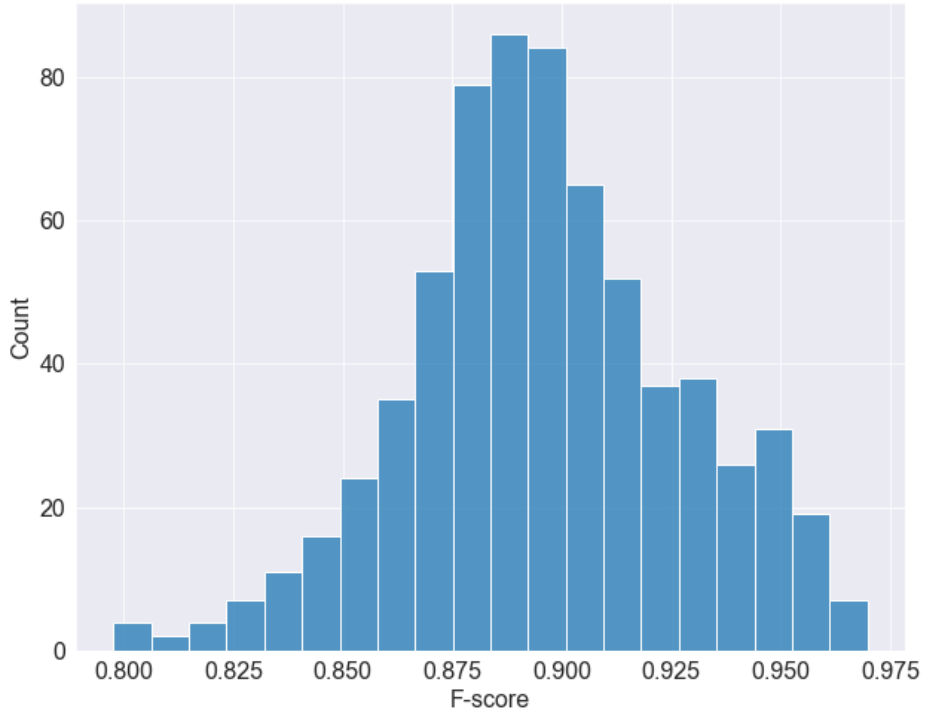
RESULTADOS

O processo de avaliação de um modelo de aprendizado de máquina envolve verificar a eficácia do modelo em prever dados não vistos anteriormente. Inicialmente, o conjunto de dados é dividido em treino e teste. Vazamento de dados refere-se à situação

em que informações do conjunto de teste são usadas no conjunto de treino, levando a modelos potencialmente superestimados. Para evitar esse problema, garantimos o uso de sinalizadores diferentes nos conjuntos de treino e teste.

Foram combinados 17 sinalizadores em grupos de três, resultando em 680 combinações distintas para grupos de teste e treino. Para cada uma das combinações, foi treinado um modelo e a avaliação de desempenho foi realizada utilizando a métrica F-score[4]. Um histograma detalhando a distribuição dos F-scores pode ser visualizado na Figura 4.

Figura 4 - Histograma dos F-scores



Fonte: Autoria própria, 2023.

CONCLUSÃO

A aplicação do aprendizado de máquina no contexto da Libras é fundamental para promover a acessibilidade e aprimorar as ferramentas de comunicação com a comunidade surda. O esforço empenhado na criação de um conjunto de dados robusto e detalhado foi fundamental para sustentar esse avanço. O modelo desenvolvido, tendo como base esse conjunto de dados, obteve uma F-score média de 89%, evidenciando sua capacidade de interpretar e traduzir sinais de Libras com alta precisão.

Futuros trabalhos têm o desafio de ampliar a capacidade do sistema para classificar sinais dinâmicos, enriquecendo ainda mais as possibilidades de interação e compreensão da Libras.

[4] O F-score é uma métrica que combina precisão e recall, com valores variando entre 0 (menos preciso) e 1 (mais preciso).

Agradecimentos

Gostaria de agradecer ao professor Gustavo Paetzold. Sua orientação foi essencial em minha trajetória. Muito obrigada por tudo.

Disponibilidade de código

Os scripts destinados à geração do dataset e ao treinamento do modelo estão disponíveis para consulta e uso no repositório[5] do Github.

Conflito de interesse

Não há conflito de interesse.

REFERÊNCIAS

CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. **Machine learning**, v. 20, p. 273-297, 1995.

CRISTIANO, A. **O que é Libras?** Disponível em: <https://www.libras.com.br/o-que-e-libras>.

LANG, S.; BLOCK, M.; ROJAS, R. Sign Language Recognition Using Kinect. **Artificial Intelligence and Soft Computing**, p. 394–402, 2012.

LUGARESI, C. et al. MediaPipe: A Framework for Building Perception Pipelines. **arXiv:1906.08172 [cs]**, 14 jun. 2019.

[5] https://github.com/luanaccampos/libras-net