

---

UNIVERSIDADE FEDERAL DE OURO PRETO  
DEPARTAMENTO DE CIÊNCIA DE COMPUTAÇÃO  
PROCESSAMENTO DIGITAL DE IMAGENS

### **Projeto de Pesquisa**

## **RECONHECIMENTO DE SINAIS DINÂMICOS DE LÍNGUA DE SINAIS**

---

### **Alunos:**

Halliday Gauss,

Thiago Borba,

Lucas de Souza.

## **Resumo**

Sabendo da importância da língua e da comunicação para as interações sociais e relações interpessoais, principalmente da Língua Brasileira de Sinais (Libras) no Brasil, e a dificuldade que as pessoas que necessitam dessas línguas para se comunicar passam, aparatos tecnológicos são relevantes para ajudá-las a se integrar na sociedade além de possibilitar para todas as pessoas uma melhor aprendizagem da Libras. Esses aparelhos podem ser aplicados em diversas áreas para produzir igualdade e inclusão, como por exemplo na saúde e educação. Este trabalho apresenta os trabalhos relacionados à área de reconhecimento de sinais em línguas de sinais juntamente do considerado estado da arte atual, assim como uma implementação de um protótipo de reconhecimento de sinais do alfabeto da Libras em vídeos utilizando Python e OpenCV. O protótipo utiliza o modelo já treinado COCO CAFFE para fazer o reconhecimento, e utiliza 22 pontos chaves considerando a mão toda como um único objeto. Para o reconhecimento é considerado a posição, altura e proximidade entre os pontos chaves. Apesar de realizar um bom reconhecimento dos sinais do alfabeto de Libras, o protótipo não reconhece muito bem os sinais representados pelas letras H, J, K, X, Y, Z, por não levar em consideração a transição dos pontos chaves.

# 1 Introdução

A comunicação, juntamente da língua, é de suma importância para que haja as interações sociais e relações interpessoais. Pessoas com deficiência, como os surdos por exemplo, também precisam dessas interações tanto com pessoas não deficientes quanto com pessoas deficientes, e para isso ocorrer são utilizadas as línguas de sinais.

As línguas de sinais ou línguas gestuais são idiomas visuais baseados nos movimentos das mãos e das expressões faciais e corporais. Essas línguas geralmente surgem nas comunidades de pessoas surdas e podem se derivar de outras línguas de sinais. Nesses tipos de línguas existem um sistema de combinação que a partir de unidades simples, formam-se unidades mais complexas. As frases e sentenças são formadas a partir de palavras que por sua vez são formadas a partir de unidades menores, morfemas, e os morfemas, são formadas a partir de queremas (menor unidade de um sinal).

Geralmente, nessas línguas, existem dois tipos de sinais: estáticos e dinâmicos. Os sinais estáticos são aqueles onde não existe um movimento para representá-los, e são representados por uma única imagem, a maioria das letras do alfabeto, por exemplo, são estáticas. Já os sinais dinâmicos são aqueles que precisam de movimento para representá-los, e são apresentados por um vídeo, as palavras nas línguas de sinais, por exemplo, são dinâmicas.

Segundo [da Silva et al. \(2017\)](#), ao longo do tempo as pessoas com deficiência, em especial os surdos, tiveram que enfrentar grandes obstáculos para serem inseridas na sociedade de forma participativa, essas pessoas, foram excluídas da sociedade e por muitos anos tiveram seus direitos básicos negados, como o de saúde, moradia e educação, sendo assim excluídas da sociedade. Portanto, assim como as línguas orais-auditivas, uma língua de sinais, que também é considerada pela linguística como língua natural, também possui grande importância para a comunicação e inclusão de pessoas na sociedade.

No Brasil, a língua de sinais adotada pela comunidade de surdos e regulamentada por lei é chamada de Libras, que é a abreviação para Língua Brasileira de Sinais. Em Libras, cada palavra é representada por um sinal, além disso essa língua tem as seguintes características:

- Configuração das Mão (CM): é a forma adotada pelas mãos, como cada letra do alfabeto.
- Ponto de Articulação (PA): é a localização das mãos em relação ao corpo.

- Movimento (M): é o modo como as mãos configuradas se movimentam ou não.
- Expressões Não Manuais (ENM): são as ações realizadas por outras partes do corpo que não as mãos – nesse caso, principalmente a expressão facial.

A Figura 1 mostra como são os sinais para representar as letras do alfabeto em Libras:

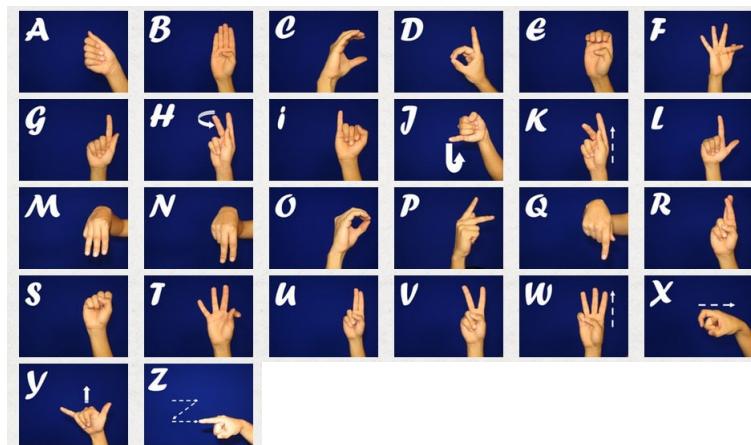


Figura 1: Sinais do Alfabeto em Libras.

A Língua Brasileira de Sinais tem muita importância no Brasil em diversas áreas. Na saúde Mendes et al. (2021) diz que "Tal necessidade se intensifica diante da tecnicidade excessiva presente no cotidiano das práticas em saúde, especialmente na área médica. É fundamental formar profissionais habilitados a compreender e a comunicarem-se adequadamente com os seus pacientes, atentos às singularidades destes.".

Na educação, Azevedo & Alencar (2021) enfatiza a importância de se aprender Libras desde a educação infantil. A mesma é de grande relevância para inclusão social e para o desenvolvimento social e emocional das pessoas surdas, essa inclusão só acontecerá de forma eficaz se existir profissionais capacitados para trabalhar de forma que todos os alunos sejam vistos dentro da sala de aula de maneira igualitária, onde incluir o ensinar seja o principal objetivo do professor. Ou seja, é imprescindível que desde o ensino infantil haja um suporte para que os alunos especiais entendam uns aos outros e aos professores, evitando problemas de aprendizagem e dificuldades de inclusão futuras.

Dado a importância da Língua Brasileira de Sinais na comunidade, a implementação de métodos tecnológicos que permitem detectar e reconhecer a linguagem de sinais em imagens e vídeos,

em tempo real ou não, teriam uma aplicação direta na sociedade, impactando de forma positiva a mesma, podendo gerar uma maior inclusão entre as pessoas, além da melhoria no aprendizado de Libras. Portanto, seria relevante a implementação de métodos que permitem o reconhecimento da Libras.

Este documento apresenta os trabalhos relacionados a área de reconhecimento de sinais em línguas de sinais juntamente do considerado estado da arte atual, assim como uma implementação de um protótipo de reconhecimento de sinais de Libras em vídeos utilizando Python e OpenCV, onde para o reconhecimento dos sinais é considerado os pontos chaves da mão e modelos já treinados. Também será mostrado os resultados obtidos pelos experimentos com o protótipo.

## 2 Revisão Bibliográfica

Nesta seção é apresentado os trabalhos relacionados com o reconhecimento de sinais dinâmicos de línguas de sinais, e o estado da arte.

### 2.1 Trabalhos Relacionados

Em Gonçalves et al. (2016) é desenvolvido um sistema que, através da captura de imagens gestuais via Kinect (sensor utilizado no videogame XBox 360 da Microsoft) e do treinamento da Rede Neural Artificial (RNA), traduz os gestos da Linguagem Brasileira de Sinais (LIBRAS). Para o desenvolvimento desse sistema foi utilizado as linguagens de programação Java e C#, nas plataformas NetBeans IDE 8.0.2 e Microsoft Visual Studio Express.

O primeiro passo dessa abordagem é rastrear o movimento da mão do usuário e salvar as imagens com o fundo removido e convertido para preto. Em seguida ocorre a conversão da imagem para escala de cinza e a aplicação do filtro laplaciano, que é utilizado para detectar e destacar as bordas.

O melhor resultado obtido com a solução desenvolvida foi de 88% de acertos utilizando uma amostra de 420 imagens, porém, ainda que o resultado tenha sido positivo, destacam-se alguns pontos a melhorar, como o reconhecimento dos gestos de forma automática e em tempo real.

Na solução proposta por Nandhini et al. (2021), utiliza-se *Convolutional Neural Network* (CNN) para classificar imagens de linguagem de sinais e um sistema de filtragem para melho-

rar a precisão do reconhecimento em áreas com iluminação variada e diferentes tons de pele. O objetivo deste trabalho é reconhecer os sinais com máxima precisão independente se está bem ou mal iluminado. Nesse processo as técnicas utilizadas são:

- Técnica de filtragem para reconhecer a área em que o sinal está contido.
- Redimensionamento da imagem, diminuindo o número de pixels.
- Conversão da imagem para uma escala de cinza, com o intuito de melhorar o desempenho, visto que imagens coloridas tem mais bits.
- Eliminação do fundo da imagem, ao redor da região da mão.
- Processo de mascaramento para ajudar na identificação das bordas, ocultando algumas partes da imagem e revelando outras.

Para execução dos processos descritos é utilizado a linguagem de programação Python; o *framework* Tensorflow para *Machine Learning* (ML); a biblioteca OpenCV, que oferece funcionalidades de visão computacional e ML; e outra biblioteca chamada NumPy que oferece suporte para matrizes multidimensionais.

Para o treinamento do modelo CNN foi utilizado um total de 1750 imagens e por fim o sistema proposto conseguiu classificar 125 palavras, a maior precisão atingida foi de 90%.

No artigo de [Amaral et al. \(2017\)](#) foi proposto um método para reconhecer gestos estáticos da mão representados por imagens de profundidade, com aplicações no desenvolvimento de interfaces naturais de usuário e reconhecimento da Língua Brasileira de Sinais. O método para resolver esse problema baseia-se na aprendizagem de máquina usando redes neurais convolucionais que tem como entrada imagens de profundidade extraídas do sensor RealSense. Ao apresentar uma imagem da mão ao sensor RealSense, é capturado uma imagem em profundidade e a mesma passa por um processo de segmentação e binarização onde é extraído somente a parte da imagem que corresponde a mão a qual é colocada em preto e branco (fundo preto e sinal da mão em branco). Em seguida essa nova imagem é esqueletonizada aplicando-se o operador de Transformada de Distancia, e por fim a imagem resultante é utilizada para treinamento e reconhecimento usando uma rede neural convulcional.

O método foi implementado utilizando a linguagem C++, OpenCV e a SDK do sensor RealSense na etapa de aquisição da imagem, e para o pré-processamento, treinamento e classificação das mesmas foi utilizado a linguagem Python com as bibliotecas Keras e Tensorflow.

Para testar e treinar a rede neural foram coletadas 1400 imagens com 14 posições distintas de mãos que correspondem a sinais de Libras, e foi alcançado uma taxa de reconhecimento de 96.42%. Porém uma desvantagem dessa proposta é que, apesar de ser afirmado que o método apresentado pode servir como base para detecção de sinais dinâmicos, não ficou claro como isso pode ser feito.

Em [Passos \(2019\)](#) é mostrado uma pesquisa na área de reconhecimento da Língua Brasileira de Sinais, juntamente de um protótipo de um software capaz de reconhecer gestos do alfabeto da língua em questão. Esse protótipo tem como objetivo reconhecer gestos em Libras presente em imagens e retornar o seu significado em forma textual, possibilitando o entendimento do que está sendo mostrado na imagem por qualquer pessoa alfabetizada.

A técnica utilizada para realizar o reconhecimento dos gestos utiliza arquiteturas de deep learning juntamente do python e o módulo de reconhecimento de objetos da biblioteca TensorFlow.

No protótipo as imagens são capturadas através de uma webcam que deve estar previamente conectada ao computador, e localizada em um ambiente de iluminação e background controlados.

Por fim, a solução proposta neste trabalho alcançou 90,28% de acurácia e 90,38% de média, com desvio padrão de 6,54%, comprovando que as técnicas de Visão Computacional permitem identificar padrões para efetuar o reconhecimento de gestos do alfabeto da Libras em imagens e vídeos. No entanto, esse trabalho não levou em consideração os gestos do alfabeto que dependem de parâmetros complexos da língua, como o movimento, para serem executados e reconhecidos.

A proposta de [Cruz et al. \(2020\)](#) sugere que já existem tecnologias de transcrição da Libras para o português e reconhecimento de gestos, e foram desenvolvidas utilizando técnicas de visão computacional e redes neurais convolutivas. Nesse artigo também é dito que essas técnicas supracitadas são consideradas estado da arte. Porém, ainda não foi encontrada uma solução efetiva para o problema, devido, ao alto custo financeiro de implantação, como por exemplo, a utilização de dispositivos específicos de aquisição de dados, equipamentos desfavoráveis, visto que algumas soluções utilizam sensores portáteis, como luvas equipadas com sensores de movimento, entre outros. Somente alguns trabalhos estão direcionados ao uso de técnicas de transferência de aprendizado e

aumento de dados, ou então a fusão de diferentes canais de dados.

Portanto, o trabalho supracitado elabora um método para reconhecimento de sinais da Libras, que seja de baixo custo, não intrusivo e eficiente no reconhecimento de sinais que são executados em movimento. Consequentemente, esse método reconhece sinais estáticos e dinâmicos da Libras, e emprega uma combinação entre rede neural convolutiva tridimensional, fusão de dados de múltiplos canais e transferência de aprendizado, por meio de um modelo de 3D-CNN. Essa técnica foi validada experimentalmente, porém foi utilizado a base de dados LIBRAS\_APOEMA gerando uma limitação no modelo, pois faz com que ele aprenda padrões específicos dessa base de dados, prejudicando o desempenho do mesmo.

O foco do trabalho [Simon et al. \(2017\)](#) é a criação de um sistema multi-câmera para treinar detectores de baixa granularidade para pontos-chave que são propensos à oclusão, como as articulações de uma mão. Denominaram o este processo descrito como bootstrapping multiview: primeiro, um ponto-chave inicial detector é usado para produzir rótulos ruidosos em múltiplas visualizações a mão. As detecções ruidosas são trianguladas em 3D usando geometria multiview ou marcados como outliers. Seguidamente, as triangulações reprojetadas são usadas como novos dados de treinamento rotulados para melhorar o detector. Esse processo é repetido várias vezes, gerando mais dados rotulados em cada iteração. Então analiza-se o resultado que relaciona analiticamente o número mínimo de visualizações para atingir as taxas alvo de verdadeiros e falsos positivos para um determinado detector. Esse método é utilizado para treinar a detecção de pontos-chaves da mão para imagens únicas, e o detector de pontos-chaves resultante é executado em tempo real em imagens RGB e tem precisão comparável aos métodos que usam sensores de profundidade. O detector de visualização única, triangulado em várias visualizações, permite a captura de movimento manual sem marcadores 3D com interações complexas de objetos.

Apesar desse método detectar muito bem sinais através da mão ele não é robusto o suficiente para trabalhar com menos câmeras e ambientes menos controlados, com muitas pessoas ou telefone celulares, por exemplo, ou ambientes mal iluminados.

No trabalho [Cheok et al. \(2019\)](#), é feita uma revisão completa sobre as técnicas de ponta utilizadas em pesquisas recentes em torno do reconhecimentos de gestos manuais e linguagem de sinais. As técnicas são categorizadas a partir das etapas de aquisição de dados, pré-processamento, segmentação, extração de características e classificação. Apesar das expressões faciais serem usadas

como parte da linguagem de sinais, tal artigo não discute acerca desse elemento.

No que diz respeito aos desafios do reconhecimento de gestos é dito que, elementos como iluminação de fundo, velocidade de movimento e a diferença do ponto de vista em ambientes 2D, podem afetar a capacidade preditiva do modelo. Portanto, existem diversos métodos de avaliação de desempenho no intuito de superar os desafios apresentados. Sendo alguns desses, a medição da escalabilidade, robustez, desempenho em tempo real e independência do usuário.

Além das dificuldades citadas, o artigo mostra também os tipos de abordagens possíveis para o reconhecimento de gestos de mão. Segundo o autor, pode-se alcançar o reconhecimento usando abordagens baseadas em visão, que requerem a aquisição de imagens ou vídeo dos gestos das mãos por meio de câmera de vídeo, ou baseadas em sensores, que requer o uso de instrumentos para capturar o movimento, a posição e a velocidade da mão.

Apesar de existirem lacunas significativas a serem preenchidas para que o reconhecimento de gestos possa ser colocado em uso, o artigo referenciado, conclui a partir de trabalhos anteriores, que HMMs (Hidden Markov Models) para determinar informações de trajetória, aparecem como abordagens promissoras para o reconhecimento dinâmico de gestos, uma vez que foi implementado com sucesso em muitas pesquisas.

Já no reconhecimento estático de gestos, o SVM (Support Vector Machine) é o método mais popular, pois tem demonstrado melhor desempenho em diversas pesquisas. Várias variantes são propostas para o método existente e híbridos de métodos estão se tornando mais amplamente utilizados, pois podem superar a deficiência do método único.

Monteiro et al. (2016) apresenta um sistema de baixo custo para reconhecimento de LIBRAS utilizando visão computacional. No artigo é apontado a falta de uma base de dados robusta para treinar os algoritmos de reconhecimento LIBRAS, o que acaba limitando o desenvolvimento de pesquisas aprofundadas sobre o tema. Por esse motivo, uma das contribuições desse trabalho foi a criação de uma base de dados com 548 vídeos, contemplando 24 palavras executadas por diversos voluntários em 3 tipos de plano de fundo. O sistema foi desenvolvido em Matlab e para reconhecer os gestos foi utilizado um método composto por três etapas:

- Obtenção da sequência residual subtraindo quadros adjacentes.
- Obtenção da matriz de características através da acumulação do movimento em células acu-

muladoras.

- Classificação usando k-Nearest Neighbors (k-NN).

Para a realização dos testes foram utilizados 450 dos vídeos presentes na base e resultou numa taxa média de acerto de 75%, o que foi considerado satisfatório, ainda que seja possível potencializar o desempenho global, melhorando o processo extração de características, bem como empregando classificadores mais robustos.

Bantupalli & Xie (2018) tem como objetivo tornar possível a comunicação entre pessoas que sabem linguagem de sinal e pessoas que não sabem. Desse modo, o artigo apresenta um sistema que usa a base de dados MNIST da *American Sign Language* (ASL) e um modelo CNN para reconhecimento dos gestos na linguagem de sinal. Inicialmente as imagens são capturadas por uma webcam, em seguida são convertidas para um escala de cinza, passam por um processo de dimensionamento e transformação e depois alimentam o modelo CNN para que ocorra a predição e classificação dos gestos e por fim sejam convertidos em texto. Para o treinamento do modelo CNN foram utilizadas 27455 amostras do conjunto de dados MNIST ASL. O conjunto de dados utilizados para teste foi de 7172 amostras e o resultado obtido foi uma taxa de acerto superior a 93%. O principal problema da solução proposta é o fundo das imagens, uma vez que o método utilizado não suporta a subtração de fundo da imagem quando os *frames* são obtidos de um vídeo.

Em Voigt (2018) é proposta uma abordagem de *Deep Learning* para reconhecimento de gestos estáticos e dinâmicos da mão, com aplicações em sinais de Libras. Através de dados capturados pelo dispositivo *Leap Motion*, incluindo tanto imagens quanto esqueletos da palma da mão, foram avaliadas diversas arquiteturas de Redes Neurais para reconhecer os gestos. As metodologias podem ser descritas em três etapas. A primeira, busca reconhecer os gestos estáticos (poses) usando redes *perceptron* multicamadas (MLP) para os dados do esqueleto, redes convolucionais (CNN) para as imagens, e redes de múltiplas entradas, utilizando ambos os tipos de informação. Na segunda, é classificado individualmente gestos que incluem movimento, e para tanto incluímos camadas recorrentes *Long Short-Term Memory* (LSTM). Para tornar o processo ainda mais preciso, foi aplicada Transferência de Aprendizado nos blocos convolucionais, trazendo os parâmetros já treinados com as poses estáticas para dentro da rede projetada para os gestos dinâmicos. Por fim, apresentam um novo algoritmo que permite reconhecer online os mesmos gestos dinâmicos da

etapa anterior, mas executados de forma sequencial, sem pausas, e sem ter informação sobre o início e final da execução de cada gesto.

### 3 Metodologia

Foi construído um protótipo baseado na ideia do artigo de Simon et al. (2017), ou seja, o protótipo é capaz de realizar a detecção de sinais de LIBRAS através de pontos chaves da mão. Para atingir tal objetivo foi utilizado um modelo já treinado chamado COCO CAFFE o qual possui um conjunto de treinamento, validação e teste da COCO (Common Objects in Context), contendo mais de 200.000 imagens e 250.000 pessoas nomeadas com pontos chaves, sendo a maioria dessas imagens em escalas médias e grandes. Esse modelo COCO é treinado pelo CAFFE Deep Learning Framework. Na figura 2 abaixo é possível ver um exemplo de algumas imagens da base de dados do modelo COCO.

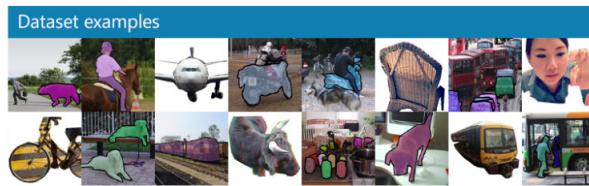


Figura 2: Exemplo de imagens da base de dados do modelo COCO.

O modelo COCO CAFFE utiliza Redes Neurais Convolucionais para o treinamento do modelo. Na primeira etapa é aplicado, pela Rede Neural, a convolução nas imagens através de um detector de características que também é gerado pela Rede, o que vai resultar numa imagem de menor tamanho que é chamada de Mapa de Características, em seguida é utilizada a função Relu ou função de ativação que vai atribuir 0 para os valores menores que 0 e vai manter os valores no Mapa de Característica para os valores maiores ou iguais a 0. A figura 3 abaixo exemplifica esse processo.

Após esse procedimento é feito um Max Pooling que retornará uma imagem menor focada nos maiores valores que são as características mais importante do Mapa de Características. Veja a figura 4.

### Etapa 1 – Operador de convolução (Relu)

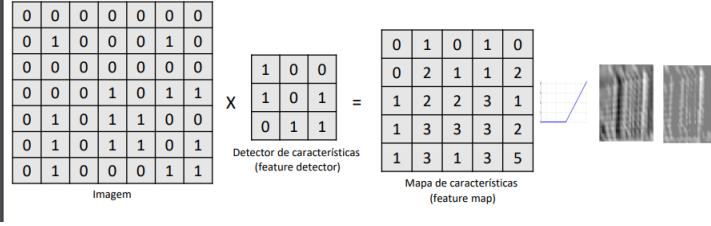


Figura 3: Processo de geração do Mapa de Características do modelo COCO CAFFE.

### Etapa 2 – Pooling

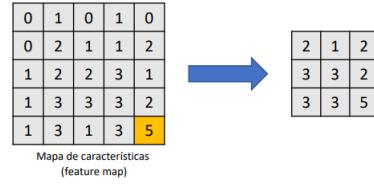


Figura 4: Aplicação do Max Pooling no Mapa de Características.

Em seguida é feito um Flattening, ou seja, uma conversão do resultado do Pooling para um vetor que é submetido para uma Rede Neural Densa Tradicional, que depois usará algum algoritmo de classificação, Soft Max por exemplo, para melhor classificação de um ponto chave. Observe a figura 5.

### Etapa 3 – Flattening

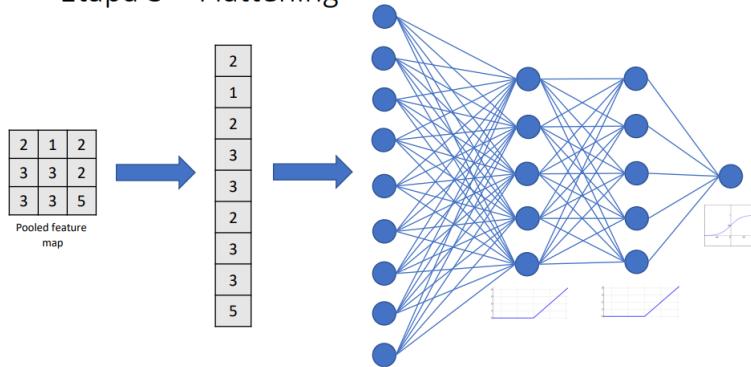


Figura 5: Aplicação do Flattening.

Cabe ressaltar que é possível, a partir de uma imagem, gerar vários Mapas de Características, assim como mostra a figura 6.

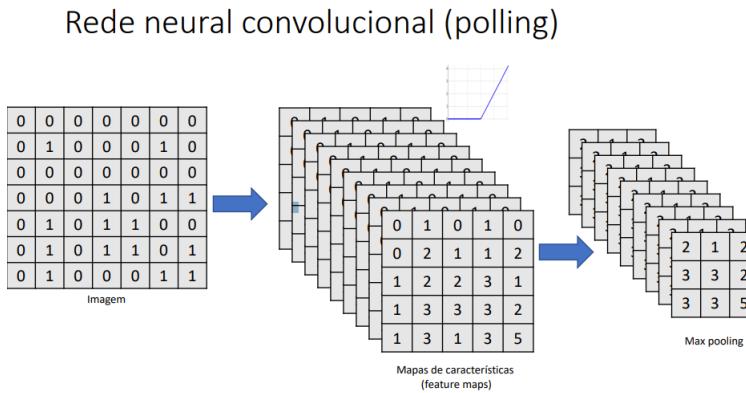


Figura 6: Geração de vários Mapas de Características.

É interessante comentar que o modelo COCO CAFE tem como base a Arquitetura VGGNET que é uma Rede Neural Convolucional, e esse modelo precisa de uma imagem de entrada, a sua largura e comprimento, e então a saída será a marcação em 2D do pontos chaves do objeto. E para cada ponto chave será retornado uma matriz de pontos candidatos. Nesse procedimento é previsto os Mapas de Confiança que verificam a probabilidade de cada ponto candidato estar na posição correta. Por exemplo, na ponta do polegar têm-se vários pontos candidatos e deve ser escolhido o ponto candidato que tem a maior confiança (probabilidade), para que então esse ponto candidato se torne um ponto chave.

Também é previsto nessa arquitetura os Mapas de Afinidade, que indicam se um ponto pode ser associado com outro. Por exemplo: o ponto do topo do polegar e o ponto um pouco abaixo do mesmo devem estar associados em uma linha reta, ou seja, estão bem próximos entre si.

Dado a explicação do funcionamento do modelo COCO CAFFE e sua arquitetura, é possível compreender melhor como foi implementado o protótipo. O protótipo baseia-se no uso do modelo COCO CAFFE para o reconhecimento dos pontos chaves da mão, o que irá permitir reconhecer sinais em Libras. Alguns detalhes sobre o modelo utilizado:

- Trata-se a mão inteira como um único objeto.
- Este modelo (COCO CAFFE) foi treinado sobre um pequeno conjunto de imagens de mão

rotuladas e usam uma Rede Neural para obter estimativas aproximadas dos pontos chave da mão.

- Os autores deste modelo possuíam um enorme sistema multi-view configurado para capturar imagens de diferentes pontos de vista ou ângulos, compostos por 31 câmeras HD.
- Os autores deste modelo usaram detectores de ponto chave e imagens de vários ângulos diferentes para criar um detector aprimorado.
- Este modelo possui 22 pontos chave. A mão tem 21 pontos, enquanto o ponto 22 é representado pelo fundo da imagem.
- A saída possui 22 matrizes, sendo cada matriz o mapa de probabilidade de um ponto chave.

A figura abaixo mostra os pontos chaves da mão que são mapeados pelo modelo:



Figura 7: Pontos chaves da mão considerados pelo modelo.

Utilizando o modelo supracitado o protótipo implementado realiza o reconhecimento, em vídeos, de todos os símbolos em LIBRAS que representam o alfabeto, com exceção das letras H, J, K, X, Y e Z. A figura abaixo mostra todas as os sinais que representam cada letra do alfabeto em Libras, e os sinais que estão riscado com um 'x' vermelho não são reconhecidos pelo protótipo.

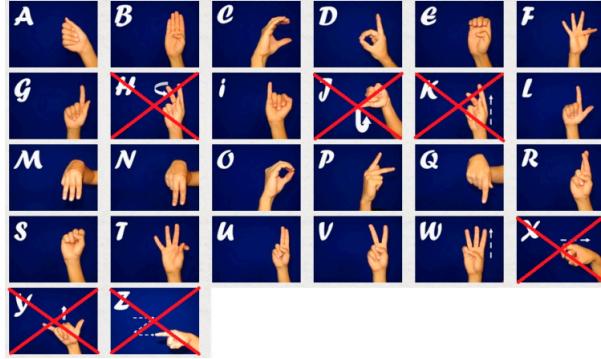


Figura 8: Sinais em Libras reconhecidos pelo modelo e pelo protótipo.

O protótipo utiliza o Google Colab, Python e OpenCV, para realizar esse reconhecimento. Alguns módulos implementados e vídeos contendo os sinais de libras são acessados pelo protótipo através do Google Drive, e então a partir de cada frame do vídeo o algoritmo tenta fazer o reconhecimento com base no modelo supracitado, e então é gerado como saída um vídeo contendo o rótulo, que é a letra representada pelo sinal, de cada sinal em cada frame analisado. Nesse vídeo também é mostrado os pontos chaves detectados nas mãos em cada frame e a ligação entre eles(esqueleto). Veja a figura 9



Figura 9: Reconhecimento da letra 'A' em Libras pelo protótipo.

Para a identificação mais dinâmica de várias posições, foram desenvolvidos 4 módulos para extrair algumas características e comparar o deslocamento das articulações (pontos chave). Foi extraído de cada frame a altura, posição e proximidade entre os pontos chaves detectados:

- Altura: verifica se um ponto está acima ou abaixo de outro ponto específico. Por exemplo, se a ponta do dedo está acima do punho.
- Posição: verifica se os dedos estão dobrados ou esticados, na posição horizontal ou na vertical. Também recebe o resultado da altura para saber em que posição a mão está (voltada

acima ou abaixo).

- Proximidade: comparar a proximidade entre os pontos chaves detectados. Por exemplo, se o resultado da Posição for igual a 'dobrado' para o dedo indicador e dedo médio e ambos estiverem na mesma altura, então significa que os dedos estão próximos.

Após extrair todas estas características, é criado o alfabeto de características, onde uma matriz de características recebe o resultado de todos os módulos extractores. Por fim, todas essas informações são utilizadas para comparar com uma nova análise (nova imagem) de entrada. A figura 10 mostra um pouco dos critérios analisados.



Figura 10: Análise da altura, proximidade e posição dos pontos chaves.

## 4 Resultados

O protótipo obteve bons resultados em reconhecer e detectar sinais do alfabeto da Libras em vídeos. A figura abaixo mostra a detecção de alguns sinais pelo protótipo:

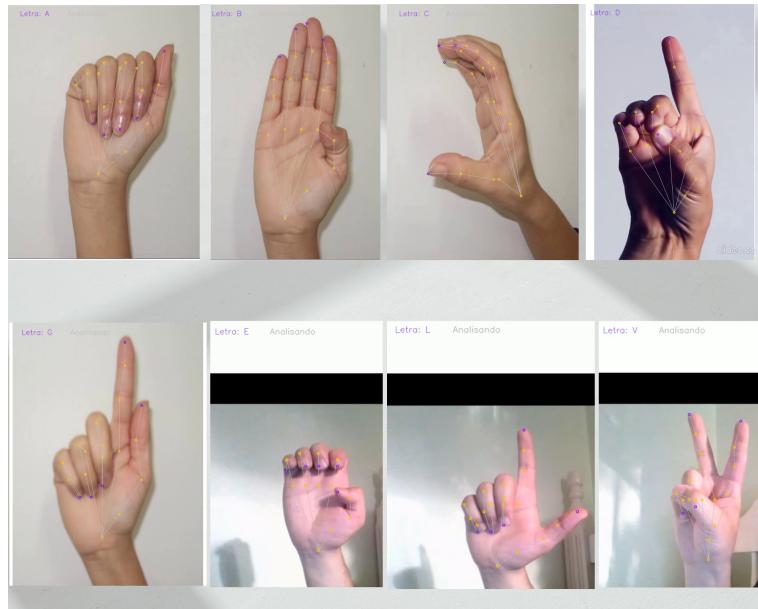


Figura 11: Reconhecimento de alguns sinais em Libras pelo Protótipo.

Apesar do bom funcionamento do protótipo, o mesmo possui algumas limitações. Devido ao movimento adicional para a execução correta das letras H, J, K, X, Y, Z elas devem ser analisadas em uma função diferente, onde comparamos a transição entre os pontos, e esse protótipo não garante o reconhecimento das mesmas. Uma outra dificuldade é a análise da letra T, para o dedo polegar, devido a estar sobreposto pelo dedo indicador, o algoritmo não reconhece os pontos da ponta dos dedos. A letra N e U podem ser confundidas no reconhecimento, pois são parecidas e mudam somente a direção.

## Referências

- Amaral, L., Lima, G., Vieira, T., & Vieira, T. (2017). Reconhecimento de gestos estáticos da mão usando a transformada de distância e aplicações em libras. *Universidade Federal de Alagoas*.
- Azevedo, L. F., & Alencar, R. M. G. (2021). A importância do ensino da língua brasileira de sinais-(libras) para educação infantil e formação dos professores das séries iniciais. *Brazilian Journal of Development*, 7(1), 5648–5671.
- Bantupalli, K., & Xie, Y. (2018). American sign language recognition using deep learning and computer vision. In *2018 ieee international conference on big data (big data)* (pp. 4896–4899).
- Cheok, M. J., Omar, Z., & Jaward, M. H. (2019, January). A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10(1), 131–153.
- Cruz, A. R. d. S., et al. (2020). Uma estratégia para reconhecimento de sinais de língua brasileira de sinais utilizando aprendizado profundo.
- da Silva, M. G. F., de Oliveira Nolasco, M., Morais, S. H. d. S. L., et al. (2017). Importância da língua materna (língua de sinais) na inclusão do aluno surdo. *Revista Includere*, 3(1).
- Gonçalves, L. C., Saad, E. F., Andrade, R. B., Romero, B. A., & de Campos, R. D. (2016). Redes neurais artificiais e processamento de imagem no reconhecimento de libras, usando o kinect. *Jornal de Engenharia, Tecnologia e Meio Ambiente-JETMA*, 1(1), 32–37.
- Mendes, V. C., de Queiroz Ribeiro, G. B. P., Lins, M. A. T., Bomfim, A. M. A., de Lima Barros, M. L. N., et al. (2021). A importância da libras na formação médica. *Pensar Acadêmico*, 19(2), 329–345.
- Monteiro, C. H. d. A., Pecoraro, L. F. I., Lacerda, A. T., Corbo, A. R., & Araujo, G. M. (2016). Um sistema de baixo custo para reconhecimento de gestos em libras utilizando visão computacional. In *Proc. 34th simpósio brasileiro de telecomunicações e processamento de sinais* (pp. 349–352).
- Nandhini, A. S., Roopan, D. S., Shiyaam, S., & Yogesh, S. (2021, may). Sign language recognition using convolutional neural network. *Journal of Physics: Conference Series*, 1916(1), 012091.

Retrieved from <https://doi.org/10.1088/1742-6596/1916/1/012091> doi: 10.1088/1742-6596/1916/1/012091

Passos, B. T. (2019). Curso de ciência da computação.

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. , 1145–1153.

Voigt, J. F. (2018). *Aprendizagem profunda para reconhecimento de gestos da mão usando imagens e esqueletos com aplicações em libras* (Master's thesis, Universidade Federal de Alagoas (UFAL), Brasil, Alagoas, Maceió). Retrieved from <http://www.repositorio.ufal.br/handle/riufal/3784>