



INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS TERESINA CENTRAL
TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

ANDRESSA RIBEIRO DE MESQUITA

**ARTICULI - PLATAFORMA DE TRADUÇÃO DE DATILOLOGIA COM BASE EM
VISÃO COMPUTACIONAL PARA FORMAÇÃO DE PALAVRAS**

TERESINA, PIAUÍ

2021

ANDRESSA RIBEIRO DE MESQUITA

ARTICULI - PLATAFORMA DE TRADUÇÃO DE DATILOLOGIA COM BASE EM VISÃO
COMPUTACIONAL PARA FORMAÇÃO DE PALAVRAS

Projeto apresentado à Banca Examinadora como requisito para aprovação na disciplina de Trabalho de Conclusão de Curso II do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí.

Orientador: Prof. Dr. Thiago Alves Elias da Silva

Coorientador: Prof. M^c. Fernando Castelo Branco Gonçalves Santana

TERESINA, PIAUÍ

2021

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD

Mesquita, Andressa Ribeiro de

M578a ArticuLi : plataforma de tradução de datilologia com base em visão computacional para formação de palavras / Andressa Ribeiro de Mesquita. - 2021.
59 f.: il. color.

Trabalho de Conclusão de Curso (Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal do Piauí, Campus Teresina Central, 2021.

Orientador : Prof. Dr. Thiago Alves Elias da Silva.

Coorientador : Prof. Me. Fernando Castelo Branco Gonçalves Santana.

1. Acessibilidade. 2. Estimativa da Pose da Mão. 3. LIBRAS. 4. Visão computacional. 5. Tradução. I. Título.

CDD - 004.21

Elaborado por Sindya Santos Melo CRB 3/1085

ANDRESSA RIBEIRO DE MESQUITA

ARTICULI - PLATAFORMA DE TRADUÇÃO DE DATILOLOGIA COM BASE EM VISÃO
COMPUTACIONAL PARA FORMAÇÃO DE PALAVRAS

Projeto apresentado à Banca Examinadora como requisito para aprovação na disciplina de Trabalho de Conclusão de Curso II do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí.

Aprovado pela banca examinadora em _____.

BANCA EXAMINADORA:

**Prof. Dr. Thiago Alves Elias da
Silva**

Instituto Federal de Educação,
Ciência e Tecnologia do Piauí - IFPI

**Prof. M^º. Fernando Castelo
Branco G. Santana**

Instituto Federal de Educação,
Ciência e Tecnologia do Piauí - IFPI

**Prof. Dr. Ricardo Martins
Ramos**

Instituto Federal de Educação,
Ciência e Tecnologia do Piauí - IFPI

**Prof. Dr. Otílio Paulo da Silva
Neto**

Instituto Federal de Educação,
Ciência e Tecnologia do Piauí - IFPI

TERESINA, PIAUÍ

2021

AGRADECIMENTOS

A caminhada que nos leva ao fim de uma jornada proporciona conhecimentos, experiências e, acima de tudo, esperança. A perseverança se reafirma à medida que abrimos a mente expandindo nossa visão, pois com isto percebemos que somos seres capazes de colaborar com a melhoria do mundo em que vivemos, seja para uma minoria ou para uma sociedade como um todo.

Sou grata àqueles que participaram dessa trajetória, que ouviram, que incentivaram, repreenderam quando necessário, mas, acima de tudo, ensinaram. Por isso, agradeço a meus pais, Vera Lúcia e Edilson Araújo, e irmãs, Thaís Ribeiro e Débora Ribeiro, por estarem presentes em uma das caminhadas mais importantes da minha vida me exortando que o estudo é um bem precioso, uma chave que abre portas de universos inimagináveis, maiores que uma sala de aula.

Agradeço aos meus amigos que se fizeram presentes e foram acolhedores constituindo uma segunda família para mim. É notável, e de suma importância dizer, que sem eles os dias não teriam a leveza de sorrisos e momentos de felicidade mesmo com tantos trabalhos e responsabilidades.

Sou grata aos docentes do curso, pois estes se empenharam a ensinar e ajudaram no meu crescimento profissional, além de deixarem uma parte inspiradora de si que foi abstraída por mim através das experiências, cada um a sua maneira.

Entre os docentes, agradeço em especial ao Prof. Dr. Thiago Alves Elias da Silva e ao Prof. M^º. Fernando Castelo Branco Gonçalves Santana por terem me ensinado e dado a oportunidade de iniciar no *Laboratory of Innovation on Multimedia Systems* (LIMS), onde pude renovar a visão de atuação do curso e participar de eventos acadêmico-científicos incentivadores.

Meu sincero obrigada!

*“Aceite o que não pode mudar, tenha a coragem de mudar o que pode
e a sabedoria para saber a diferença.”
(The Flash)*

RESUMO

Este trabalho apresenta o desenvolvimento de uma plataforma chamada ArticuLi, uma aplicação *web* responsiva que realiza a formação de palavras a partir de imagens de sinais do alfabeto datilológico da Língua Brasileira de Sinais (LIBRAS), traduzidos em sequência e com opção de reprodução em áudio. A ArticuLi utiliza um modelo de Rede Neural Convolucional pré-treinado com base em um conjunto de dados, o *Common Objects in Context* (COCO). Através do modelo são identificados *key points*, ou seja, pontos-chave que representam as articulações da mão e que são necessários para estimar a sua pose. Com uso da biblioteca OpenCV, o algoritmo analisa as coordenadas dos pontos mapeados para descobrir a distância, altura e proximidade entre eles e chegar a tradução dos sinais sequencialmente para retornar uma palavra. Os testes com o Mínimo Produto Viável (MVP) foram realizados com imagens fotografadas por 10 voluntários usando seus *smartphones* como dispositivo de captura. Com isso, obteve-se um conjunto balanceado com 5 amostras para cada sinal totalizando 130 imagens. Considerando o número total de imagens e a quantidade de sinais traduzidos corretamente, incluindo sinais estáticos e dinâmicos, a plataforma alcançou acurácia de 94,6%, ou seja, das 130 imagens analisadas, 123 foram traduzidas corretamente onde as letras A, B, C, D, E, G, I, K, L, M, N, O, P, Q, R, S, T, U, V, W e Y obtiveram 100% de acurácia, as letras F, H, J e Z 80% e X 40% de acurácia, tornando-a apta a formar palavras.

Palavras-chaves: Acessibilidade; Estimativa da Pose da Mão; LIBRAS; Visão Computacional; Tradução.

ABSTRACT

This work presents the development of a platform proposal called ArticuLi, a responsive web application that performs the formation of words from images of signs from the dactilological alphabet of the Brazilian Sign Language (LIBRAS), translated in sequence and with the option of reproduction in audio. ArticuLi uses a pre-trained Convolutional Neural Network model based on a dataset, the Common Objects in Context (COCO). Through the model, key points are identified, that is, key points that represent the hand articulations and that are necessary to estimate your pose. Using the OpenCV library, the algorithm analyzes the coordinates of the mapped points to discover the distance, height and proximity between them and translate the signals sequentially to return a word. The Minimum Viable Product (MVP) tests were performed with images photographed by 10 volunteers using their smartphones as a capture device. Thus, a balanced set with 5 samples for each signal was obtained, totaling 130 images. Considering the total number of images and the amount of correctly translated signals, including static and dynamic signals, the proposal achieved an accuracy of 94.6%, that is, of the 130 images analyzed, 123 were correctly translated where the letters A, B, C, D, E, G, I, K, L, M, N, O, P, Q, R, S, T, U, V, W and Y obtained 100% accuracy, the letters F, H, J and Z 80% and X 40% accuracy, making it able to form words.

Key-words: Accessibility; Hand Pose Estimation; LIBRAS; Computer Vision; Translation.

LISTA DE ILUSTRAÇÕES

Figura 1 – Alfabeto datilológico	19
Figura 2 – Modelo cinemático ou baseado em esqueleto	20
Figura 3 – Exemplo de modelo cinemático produzido utilizando o Human Pose Dataset MPII.	21
Figura 4 – Exemplo de saída da rede de (NEWELL; YANG; DENG, 2016) na primeira imagem (modelo cinemático) e, na sequência, mapas de calor indicando pontos de articulação localizados (pescoço, cotovelo esquerdo, pulso esquerdo, joelho direito, tornozelo direito)	21
Figura 5 – Representação dos mapas de calor para identificação de pontos chaves da mão direita.	22
Figura 6 – Demonstração dos mapas de calor com modelo da COCO utilizado na ArticuLi - configuração do sinal da letra B.	22
Figura 7 – Multiplicação das regiões da matriz da imagem de entrada pela matriz de kernel definida pela rede neural para gerar o mapa de características da camada de convolução.	23
Figura 8 – Aplicação da função <i>max pooling</i> sobre o mapa de características produzido na etapa anterior (convolução) para gerar uma submatriz menor com características relevantes.	24
Figura 9 – Representação visual do processo de extração de características e classificação realizado pela CNN. Na extração de características (tracejado vermelho) há as etapas de convolução e <i>pooling</i> em alternância, enquanto na classificação (tracejado azul) há a camada totalmente conectada.	24
Figura 10 – Amostras de categorias rotuladas do conjunto de dados do MS COCO	25
Figura 11 – <i>Interface</i> responsiva da ArticuLi em tela de dispositivo <i>mobile</i> - tradução da palavra casa: (a) página principal da plataforma, (b) tela de ação, (c) página principal da plataforma com fotografias carregadas a partir do servidor e (d) página de resultado com tradução e opção de reprodução em áudio	27
Figura 12 – Leap Motion Controller e coordenadas cartesianas X, Y e Z utilizadas no estudo.	29
Figura 13 – Logo da plataforma ArticuLi	33
Figura 14 – Diagrama de Atividade representando as ações do fluxo principal Realizar Tradução e os autores que o executam	34
Figura 15 – Primeira etapa da interação do usuário em sua jornada - Upload de foto	35
Figura 16 – Segunda etapa da interação do usuário em sua jornada - Tradução	36
Figura 17 – Terceira etapa da interação do usuário em sua jornada - Resultado	37

Figura 18 – <i>Service Blueprint</i> completo - etapas de <i>upload</i> da foto, tradução e resultado em sequência.	38
Figura 19 – Anatomia da mão - localização das articulações CMC, MCF, IFP e IFD. . .	39
Figura 20 – Representação dos pontos de articulações da mão que são mapeados pelo modelo da COCO na abordagem da estimativa da pose da mão.	40
Figura 21 – Representação dos intervalos dos dedos da mão	40
Figura 22 – Mapeamento de <i>key points</i> da mão em configuração do sinal B	41
Figura 23 – Esqueleto formado com a ligação das coordenadas de cada <i>key point</i> da mão executando sinal de B	42
Figura 24 – Resultado do processo de identificação do sinal da letra B	42
Figura 25 – À esquerda um exemplo do plano cartesiano com o sinal da letra B e pontos mapeados. À direita estão as coordenadas X e Y desses pontos.	43
Figura 26 – Exemplo prático dos dados usados no módulo posição - Dedo polegar . . .	44
Figura 27 – Fluxo das fotos na plataforma Articuli	44
Figura 28 – Ilustração de gráfico com a representação da quantidade de amostras de sinais traduzidos corretamente pela Articuli.	46
Figura 29 – Representação da (a) mão em configuração de F e (b) mão em configuração de T	46
Figura 30 – Mão em configuração de F com ligação incorreta de pontos nos dedos polegar e indicador a esquerda com círculo expondo a região de mapeamento incorreto. À direita resultado obtido.	47
Figura 31 – Mapeamento dos pontos do dedo polegar na amostra não identificada do sinal H.	48
Figura 32 – Representação do (a) esqueleto do sinal da letra I e (b) esqueleto do sinal da letra J.	49
Figura 33 – Resultado da amostra do sinal da letra J como letra I.	49
Figura 34 – Amostra não identificada com mão em configuração de Z com detalhes para o ponto do pulso e ponto 20 de extremidade do dedo mínimo	50
Figura 35 – Mãos em configuração de Z apresentadas em sequência: demonstração de momentos do movimento realizado na execução do sinal e sua tradução no canto superior esquerdo da fotografia.	50
Figura 36 – Amostras da letra X com tradução inadequada	51
Figura 37 – Características dos sinais das letras A e X pré-definidas no algoritmo. . . .	51
Figura 38 – Representação de cálculo com coordenadas do eixo Y da ponta do dedo médio e indicador.	52
Figura 39 – Representação comparativa entre os trabalhos relacionados e a plataforma Articuli detalhando dispositivos de captura, classificadores, tipo de <i>dataset</i> e quantidade de amostras	53

Figura 40 – Comparativo de características da metodologia proposta e trabalhos relacionados para evidenciar diferenciais	54
--	----

LISTA DE TABELAS

Tabela 1 – Graus da perda auditiva	17
Tabela 2 – Intervalos de identificação de cada dedo da mão	40

LISTA DE ABREVIATURAS E SIGLAS

API	<i>Aplication Programming Interface</i>
CNN	<i>Convolutional Neural Network</i>
IBGE	Instituto Brasileiro de Geografia e Estatística
IDE	<i>Integrated Development Environment</i>
LIBRAS	Língua Brasileira de Sinais
LMC	<i>Leap Motion Controller</i>
MVP	<i>Minimum Viable Product</i>
ONU	Organização das Nações Unidas
PDI	Processamento Digital de Imagens
UML	<i>Unified Modeling Language</i>
WHO	<i>World Health Organization</i>

SUMÁRIO

1	INTRODUÇÃO	14
1.1	Justificativa	15
1.2	Objetivos	15
1.2.1	Objetivo Geral	15
1.2.2	Objetivos Específicos	15
2	FUNDAMENTAÇÃO TEÓRICA	17
2.1	Deficiência Auditiva	17
2.2	Língua Brasileira de Sinais	18
2.3	Visão Computacional	19
2.3.1	Estimativa da Pose Humana	20
2.3.1.1	Pose da Mão	21
2.3.2	Rede Neural Convolucional	22
2.3.3	<i>Microsoft Common Objects in Context - COCO</i>	25
2.3.4	Ferramentas <i>web</i>	26
3	TRABALHOS RELACIONADOS	28
3.1	Tradução de Libras por Imagem	28
3.2	SingApp: Um modelo de identificação de língua de sinais através de captura de movimento em tempo real	29
3.3	Classificação Automática de Sinais Visuais da Língua Brasileira de Sinais representados por Caracterização Espaço-temporal	30
3.4	Sistema Embarcado Linux para Reconhecimento de Vogais em Libras utilizando Aprendizagem de Máquina	31
3.5	Um Módulo de Software para o Reconhecimento de Sinais Estáticos da Língua Brasileira de Sinais	31
4	ARTICULI - PROPOSTA DE PLATAFORMA DE TRADUÇÃO DE DATILOLOGIA COM BASE EM VISÃO COMPUTACIONAL PARA FORMAÇÃO DE PALAVRAS	33
4.1	Modelagem da Plataforma	33
4.1.1	Diagrama de Atividade	34
4.1.2	<i>Service Blueprint</i>	35
4.2	Implementação	38
4.2.1	Visão anatômica da mão humana	39
4.2.2	Mapeamento de <i>key points</i> com OpenCV	41

4.2.3	Análise e definição de características	43
4.2.4	Formação da palavra	44
5	RESULTADOS E DISCUSSÕES	45
5.1	Construção do banco de imagens	45
5.2	Tradução dos sinais	45
5.3	Comparativo com os Trabalhos Relacionados	52
6	CONCLUSÃO	55
6.1	Trabalhos Futuros	55
	REFERÊNCIAS	56

1 INTRODUÇÃO

Segundo dados divulgados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em sua última pesquisa, censo de 2010, mais de 45 milhões de brasileiros apresentavam algum tipo de deficiência. Dentre eles, cerca de 5% representavam pessoas com deficiência auditiva (IBGE, 2010). Em 2019, a *World Health Organization* (WHO) divulgou que mais de cinco por cento da população mundial sofria de perda auditiva incapacitante contabilizando cerca de 466 milhões de pessoas (WHO, 2019).

Diante de números que tendem a aumentar e da necessidade de comunicação e inclusão para vivência saudável em sociedade, os brasileiros portadores de deficiência auditiva tiveram sua língua própria oficializada através da Lei nº 10.436 em 24 de abril de 2002. Ela caracteriza a Língua Brasileira de Sinais (LIBRAS) como uma forma de compreender o universo visual-motor alinhado a singularidade de seu sistema gramatical e que atua como a principal ferramenta para transmissão de ideias e comunicação das pessoas surdas (BRASIL, 2002).

Apesar da existência de uma língua oficial, essas pessoas ainda convivem com barreiras em seu processo comunicativo devido a falta de domínio da libras por ouvintes (LEAL, 2018). Nesse sentido, um levantamento feito de 2010 a 2018 elenca aplicações criadas para traduzir letras ou palavras do português brasileiro para a libras com foco em aspectos computacionais e linguístico-comunicacionais através de avatares 3D (CORRÊA et al., 2018).

Nota-se, ainda, um crescimento nos estudos de tecnologias que possibilitem o inverso da tradução já conhecida, pois ainda há escassez de aplicações que realizem o processo de tradução da LIBRAS para o Português. Atualmente, as abordagens utilizadas nos estudos de reconhecimento por computador compreendem luvas com dispositivos (sensores) agregados para detecção e rastreamento de movimentos, ou técnicas de visão computacional não invasivas (SANTOS et al., 2020).

Há trabalhos de visão computacional que abordam a estimativa da pose humana através do treinamento de redes neurais. Esse método realiza a análise e o reconhecimento de articulações para mapear os *keypoints* do corpo em imagens podendo também ser aplicado a mão (RODRIGUES et al., 2018). Para isso, os modelos utilizados são treinados com uma base robusta de dados e de boa qualidade para reduzir a taxa de erro e melhorar o desempenho de suas predições (ROSTIANINGSIH; SETIAWAN; HALIM, 2020).

Diante da necessidade da existência de aplicações que gerem alguma facilidade para ouvintes e portadores de deficiência auditiva ou diminuam ruídos no ato de se comunicar, este estudo discorre sobre uma plataforma com abordagem não invasiva que utiliza a estimativa da pose da mão humana com visão computacional para identificar letras do alfabeto da LIBRAS (datilologia) em imagens e retornar a palavra formada pelo conjunto identificado.

1.1 JUSTIFICATIVA

As desigualdades sociais que acometem portadores de deficiência auditiva se concretizam através das diversas formas de exclusão ocasionadas por preconceito e desconhecimento da condição biológica e identitária desse grupo pela sociedade. Os portadores de deficiência auditiva representam uma quantidade grandiosa de pessoas que se encontram desassistidas ao longo de seu desenvolvimento humano e nos seus direitos individuais e sociais (WETTERICH; BARROSO; FREITAS, 2020).

De acordo com a WHO, estima-se que uma em cada dez pessoas até 2050 sofrerão de perda auditiva incapacitante, o equivalente a mais de 900 milhões de pessoas com perda auditiva maior que 40dB no ouvido com melhor audição em adultos e maior que 30dB no ouvido com melhor audição em crianças (WHO, 2019). Segundo a mesma organização, as consequências sociais e emocionais ocasionadas por problemas de comunicação tem impacto direto no cotidiano dessas pessoas, pois acabam criando sentimento de solidão e frustração.

A tecnologia oferece a possibilidade de criar *softwares* que podem transformar a vida das pessoas, facilitar sua realidade ou intervir de forma positiva no seu dia a dia. Por isso, com foco voltado para as dificuldades de pessoas com deficiência auditiva considerando a comunicação, este trabalho visa apresentar o desenvolvimento da plataforma proposta para contribuir no processo de comunicação, fornecendo a tradução de sinais e viabilizando a troca de palavras com ouvintes.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Apresentar uma plataforma *web* de tradução de datilologia que atenuie os ruídos sofridos durante a interação entre pessoas com deficiência auditiva e ouvintes. O desenvolvimento da plataforma almeja proporcionar a clareza de informações transmitidas pelo público deficiente auditivo.

1.2.2 Objetivos Específicos

- Estimular a interação entre portadores de deficiência auditiva e ouvintes atuando como meio facilitador;
- Promover a clareza de informações através das palavras formadas pelos sinais traduzidos;
- Disponibilizar transmissão em áudio para viabilizar o repasse da tradução de forma rápida.

DESCRIÇÃO DA MONOGRAFIA

Este capítulo apresentou a contextualização da problemática que a plataforma abrange, a justificativa do trabalho e seus objetivos. De forma sucinta, trata-se de uma aplicação que visa estimular a interação entre deficientes auditivos e ouvintes e chamar a atenção da academia para produções voltadas à comunidade surda.

Posteriormente será visto:

- **Capítulo 2 - Fundamentação Teórica:** Neste capítulo será apresentado o embasamento teórico que direcionou o projeto e serviu como base para que as etapas de desenvolvimento fossem realizadas;
- **Capítulo 3 - Trabalhos Relacionados:** Este capítulo descreve trabalhos que possuem finalidade convergente à ArticuLi ressaltando características que os diferem.
- **Capítulo 4 - ArticuLi - Plataforma de Tradução de Datilologia com Base em Visão Computacional para Formação de Palavras:** Este capítulo traz detalhes referentes ao processo de implementação da plataforma deixando claro suas etapas desde a modelagem até a implementação.
- **Capítulo 5 - Resultados e Discussões:** Este capítulo relata os testes realizados para validar a metodologia apresentada no capítulo anterior.
- **Capítulo 6 - Conclusão:** Retrata a análise do que foi desenvolvido em relação aos objetivos do projeto explicitados no Capítulo 1.

2 FUNDAMENTAÇÃO TEÓRICA

Para melhor compreensão do projeto, a fundamentação teórica elenca os conhecimentos que direcionaram a construção da aplicação fornecendo a base necessária para as suas etapas.

2.1 DEFICIÊNCIA AUDITIVA

A perda de audição nas fases iniciais da vida e seu reconhecimento tardio influenciam diretamente o desenvolvimento da fala, linguagem, cognição e socialização do indivíduo. Isso provoca limitações de atividades, interferência nas habilidades de compreensão dos discursos em ambientes silenciosos e com ruídos, restrição de participação em eventos sociais e redução da qualidade de vida nas diferentes faixas etárias (JARDIM; MACIEL; LEMOS, 2017).

Nesse contexto, a estimulação precoce das vias auditivas periféricas e centrais proporcionam impacto fundamental no desenvolvimento das habilidades auditivas e de comunicação oral possibilitando visualizar um prognóstico positivo da reabilitação auditiva. Políticas públicas e os avanços tecnológicos tem transformado o panorama da qualidade de vida de crianças através de programas de intervenção fonoaudiológica que atuam como medida preventiva às evoluções secundárias da deficiência auditiva (BICAS; GUIJO; DELGADO-PINHEIRO, 2017).

A perda auditiva é conceitualmente classificada em unilateral, bilateral, parcial ou total de 41 decibéis (dB) ou mais. O seu grau pode ser discernido em audição normal (até 25 dB), leve (de 26 a 40 dB), moderado (de 41 a 55 dB), moderadamente severo (de 56 a 70 dB), severo (de 71 a 90 dB) e profundo (maior ou igual a 91 dB). A Tabela 1 mostra cada grau mencionado e seus respectivos intervalos em decibéis. Esses dados são obtidos a partir de avaliação audiológica cujo objetivo é determinar a integridade do sistema auditivo, identificar tipo, grau e configuração da perda auditiva em cada orelha (BARBOSA et al., 2018).

Tabela 1 – Graus da perda auditiva

Grau	Decibéis (dB)
Normal	≤ 25
Leve	de 26 a 40
Moderado	de 41 a 55
Moderadamente severo	de 56 a 70
Severo	de 71 a 90
Profundo	≥ 91

FONTE: Elaborada pela autora

Na tecnologia, os Serviços de Atenção à Saúde Auditiva visam a adaptação de aparelhos de amplificação sonora individuais (AASI) como ação de fluxo contínuo e acompanhamento médico para realização de ajustes e verificação periódica das condições técnicas e resultados obtidos (JARDIM; MACIEL; LEMOS, 2017). Com isso, o aparelho funciona como recurso

terapêutico e faz com que o usuário utilize sua audição remanescente (FONSÊCA; DUTRA; FERREIRA, 2020). Sendo assim, considera-se que uma pessoa com perda auditiva interage em sociedade por meio de experiências visuais manifestando sua cultura através do uso da Língua Brasileira de Sinais (BRASIL, 2002).

2.2 LÍNGUA BRASILEIRA DE SINAIS

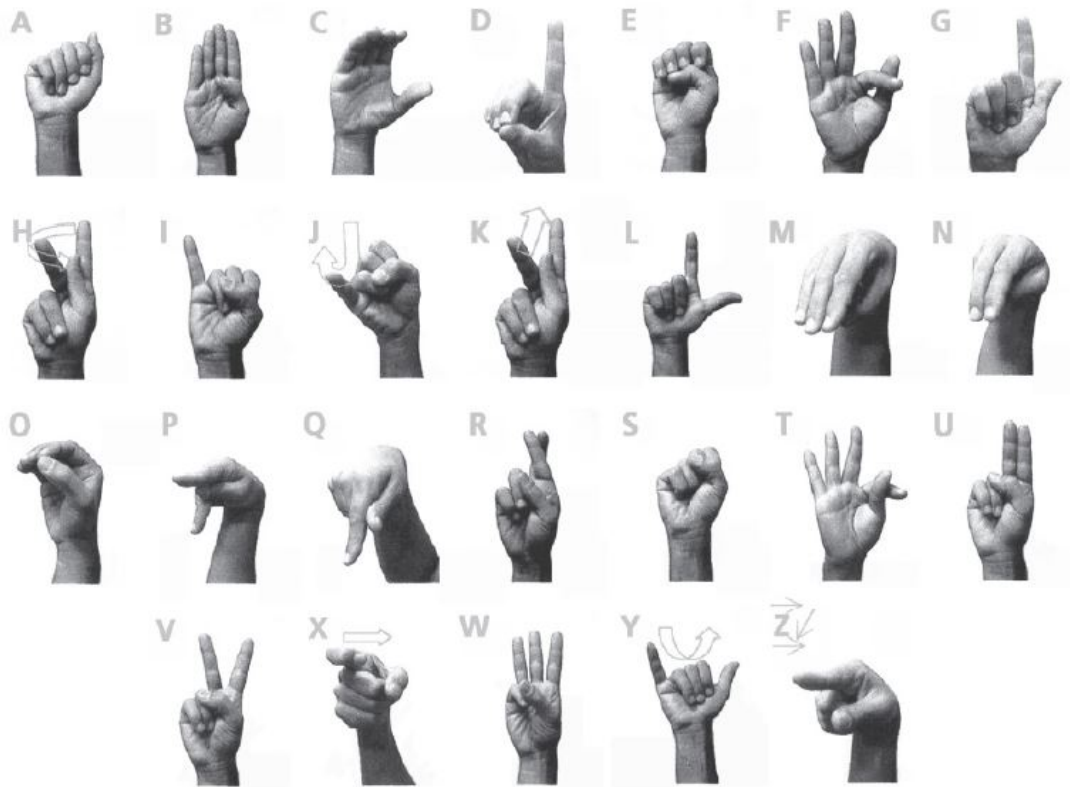
A LIBRAS é uma língua oriunda da variação de contextos, épocas e influências que está em constante processo adaptativo conforme o surgimento de mudanças. Ela é definida como natural por possuir estrutura gramatical própria analisada à nível de fonologia, morfologia e sintaxe (RODRIGUES; ALMEIDA-SILVA, 2017) atendendo a um sistema linguístico de natureza visual-motora rico capaz de atuar como forma de comunicação e expressão para que seu público possa transmitir ideias e fatos (BRASIL, 2002).

Nesse contexto, a transmissão se dá através de manifestações corporais, mais propriamente dito gestos. Com isso, o reconhecimento almejado pelos estudos da área da computação envolve a classificação de gestos considerados estáticos, trajetórios ou dinâmicos. Gestos estáticos possuem o posicionamento fixo dos dedos em relação a mão e representam informações simples como caracteres, gestos trajetórios indicam uma sequência de gestos com início e fim previamente definidos enquanto que gestos dinâmicos caracterizam-se pela variação do posicionamento dos dedos e da mão e representam informações mais complexas como palavras (BONATTO, 2018).

Cada gesto realizado com a mão é executado com forma ou configuração diferente de acordo com o sinal que representa. Em comparação à Língua Portuguesa, as configurações dos dedos e das mãos representam as emissões sonoras da oralização e compreendem a parte fonológica da LIBRAS em conjunto com outros parâmetros que ajudam na interpretação da mensagem, contextualização para o significado do sinal ou distinções do que está sendo enunciado (BENASSI; PADILHA, 2016; ROSA et al., 2016).

Levando em consideração o alfabeto datilológico, temos configurações estáticas ou dinâmicas que variam de acordo com a letra que se quer expressar conforme mostra a Figura 1 (FERNANDES; ROMEIRO, 2017).

Figura 1 – Alfabeto datilológico



FONTE: (BONATTO, 2018).

Nota-se que as letras H, J, K, X, Y e Z atendem à qualidade de sinais dinâmicos, pois há a variação da posição dos dedos e da mão com a adição de movimentos (indicados pelas setas na Figura 1). Sabendo disso, os estudos que trabalham sinais dinâmicos na computação, mais especificamente na área de Visão Computacional, utilizam vídeos para efetuar a extração de dados relevantes enquanto estudos que trabalham sinais estáticos utilizam imagens por não existir variações durante as suas execuções (MACHADO et al., 2018).

2.3 VISÃO COMPUTACIONAL

Seguindo o modelo visual humano, a Visão Computacional busca compreender os objetos variados que podem estar presentes em cenas ou contextos. De maneira mais clara, após uma imagem ser visualizada pelo olho humano, o cérebro começa a interpretá-la procurando por padrões que ajudem a deduzir sobre o que aquela imagem se refere, da mesma forma, na computação temos uma imagem sendo captada por uma câmera ou *webcam* ligada a um computador que realiza um processamento através de um algoritmo para identificar padrões e classificar o objeto que aparece na imagem (PINHEIRO; SILVA, 2018).

Do ponto de vista técnico, a Visão Computacional realiza a extração de informações de uma imagem, além de identificar e classificar objetos quando em conjunto com Inteligência Artificial. Isso é possível devido a um modelo treinado com um conjunto de imagens relacionadas

ao objeto que se quer classificar. Esse modelo funciona como o cérebro da visão computacional dando poder de interpretação a um computador (FERREIRA, 2020).

Nesse contexto, as imagens captadas são entendidas pelo computador como uma matriz bidimensional de pixels, cada um com um valor e local específico conforme suas coordenadas no plano cartesiano representado pela imagem. Esses pixels são tratados matematicamente para que o computador possa manipular e extrair seus valores numéricos e conseguir realizar o processamento (SILVA, 2020).

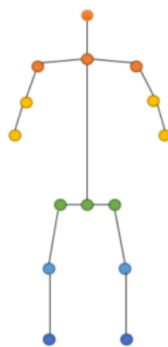
Portanto, sistemas desenvolvidos com base em visão computacional podem abordar aplicações variadas de acordo com o modelo que proporciona a sua interpretação. Alguns exemplos são reconhecimento de pessoas e de sua postura, assinaturas, objetos do cotidiano, inspeção de peças em robôs e em indústrias automatizadas, entre outros (FERREIRA, 2020).

2.3.1 Estimativa da Pose Humana

A estimativa da pose humana é um problema da visão computacional com grau de dificuldade elevado (DAI et al., 2020). Ela abrange identificar os encontros de partes do corpo humano a partir de dados de entrada, como imagens e vídeos, com uso de *neural networks*. Entre elas, o uso de *Convolutional Neural Network*, aplicado em diversos estudos da área, apresentou desempenho significativo na tarefa de classificação e por isso é comumente usado (ZHENG et al., 2021).

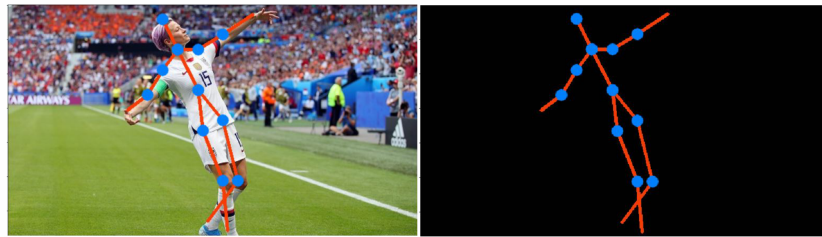
A modelagem do corpo humano é um aspecto importante da estimativa da pose humana para representar pontos-chave e recursos extraídos das imagens de entrada. Usa-se o modelo cinemático, também chamado de modelo baseado em esqueleto, conforme mostrado na Figura 2 e 3, que inclui um conjunto de posições articulares e as orientações dos membros para representar a estrutura do corpo humano (ZHENG et al., 2021).

Figura 2 – Modelo cinemático ou baseado em esqueleto



FONTE: Adaptado de (ZHENG et al., 2021).

Figura 3 – Exemplo de modelo cinemático produzido utilizando o Human Pose Dataset MPII.

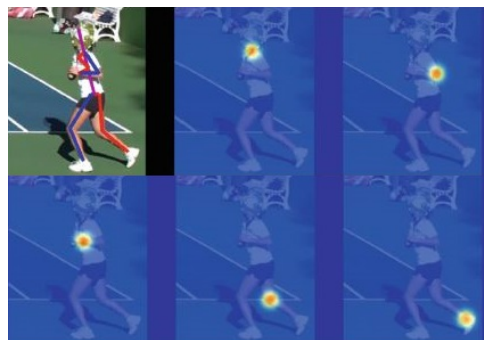


FONTE: Elaborado pela autora.

Para inferências sobre os encontros ou articulações, a abordagem que tem sido mais utilizada é a de mapa de calor, a localização precisa de pontos-chave na imagem é identificada a partir de previsões por pixel. O mapa de calor, como demonstrado na Figura 4, é utilizado para representar as coordenadas x e y que correspondem aos pontos convencionados no modelo de treinamento para cada articulação (SUN et al., 2019).

Com essa abstração, a partir do conhecimento da pose, é possível identificar pessoas em imagens e vídeos, suas ações ou qual postura adotam em determinadas cenas. Por isso, essas inferências aumentam a probabilidade de criação de aplicações variadas e necessárias a contextos específicos (NEWELL; YANG; DENG, 2016).

Figura 4 – Exemplo de saída da rede de (NEWELL; YANG; DENG, 2016) na primeira imagem (modelo cinemático) e, na sequência, mapas de calor indicando pontos de articulação localizados (pescoço, cotovelo esquerdo, pulso esquerdo, joelho direito, tornozelo direito)

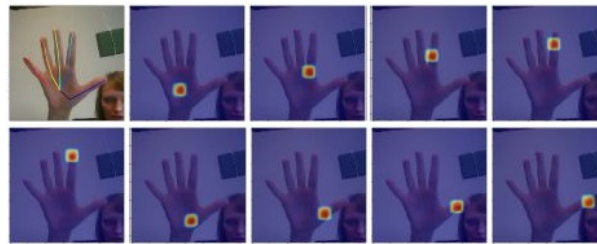


FONTE: Adaptado de (NEWELL; YANG; DENG, 2016).

2.3.1.1 Pose da Mão

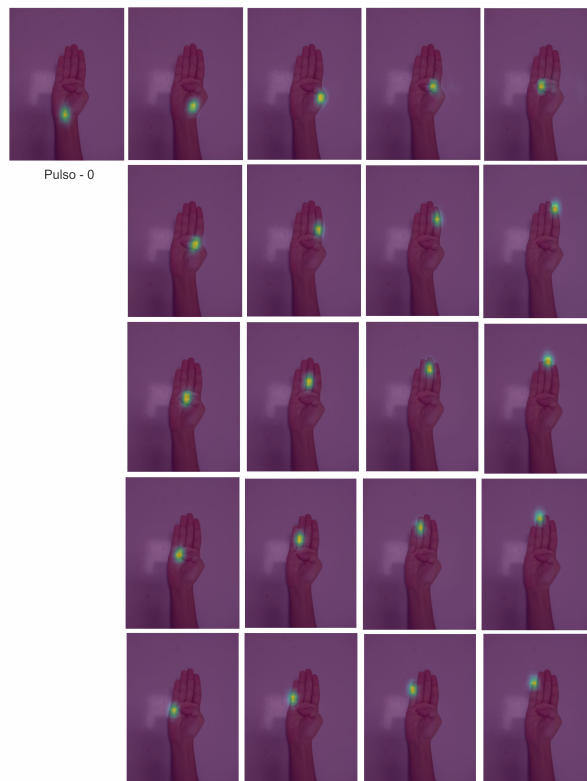
Segundo (IQBAL et al., 2018), a estimativa de um membro específico pode levar em consideração a mesma abordagem aplicada ao corpo. Nesse sentido, os autores falam que os métodos utilizados para a estimativa do corpo humano podem ser adaptados e executados para estimar a pose da mão de uma pessoa. Nas Figuras 5 e 6, temos mapas de calor, antes aplicados em imagem do corpo todo, agora aplicados às mãos.

Figura 5 – Representação dos mapas de calor para identificação de pontos chaves da mão direita.



FONTE: Adaptado de (IQBAL et al., 2018).

Figura 6 – Demonstração dos mapas de calor com modelo da COCO utilizado na ArticuLi - configuração do sinal da letra B.



FONTE: Elaborado pela autora.

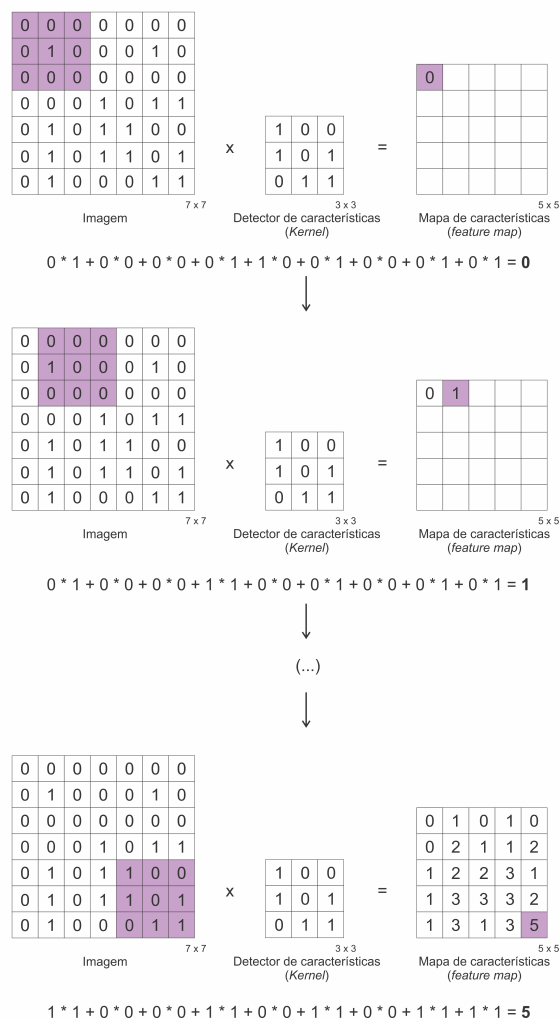
2.3.2 Rede Neural Convolucional

Como mencionado anteriormente, a *Convolutional Neural Network* (CNN) ou, em português, Rede Neural Convolucional, tem apresentado bom desempenho na tarefa de classificação (ZHENG et al., 2021). Essa rede envolve múltiplas camadas com diferentes funções que se alternam entre camadas de convolução e de *pooling* (SANTOS et al., 2019). Com ela, a matriz da imagem de entrada avança as camadas para receber a ação de seus filtros, conhecidos como *kernels* (pequenas matrizes) usadas para extrair características locais (*local features*) da matriz de entrada (COCOLO, 2020).

A primeira camada é a de convolução. Conforme demonstrado na Figura 7, seu objetivo é gerar um mapa de características (*feature map*) a partir da soma da multiplicação entre os elementos das submatrizes que formam a matriz de entrada e os elementos da matriz de *kernel* definida na rede (COCOLO, 2020).

A matriz de *kernel* é um filtro detector de características aplicado às regiões (submatrizes) da imagem de entrada (ROSA et al., 2018). Dessa forma, na Figura 7, as regiões são representadas pelo conjunto de quadrados de cor roxa e tem seus elementos multiplicados linha a linha com a matriz de *kernel*, ou seja, o primeiro elemento da região, o 0, é multiplicado pelo primeiro elemento do filtro, o 1, e o resultado dessa multiplicação deve ser somado ao resultado das multiplicações posteriores até que todos os elementos da região passem por essa operação e um resultado seja obtido para compor o primeiro elemento do mapa de características. Da mesma forma, a soma das multiplicações é repetida até que todas as regiões sejam percorridas e o mapa de características seja finalizado.

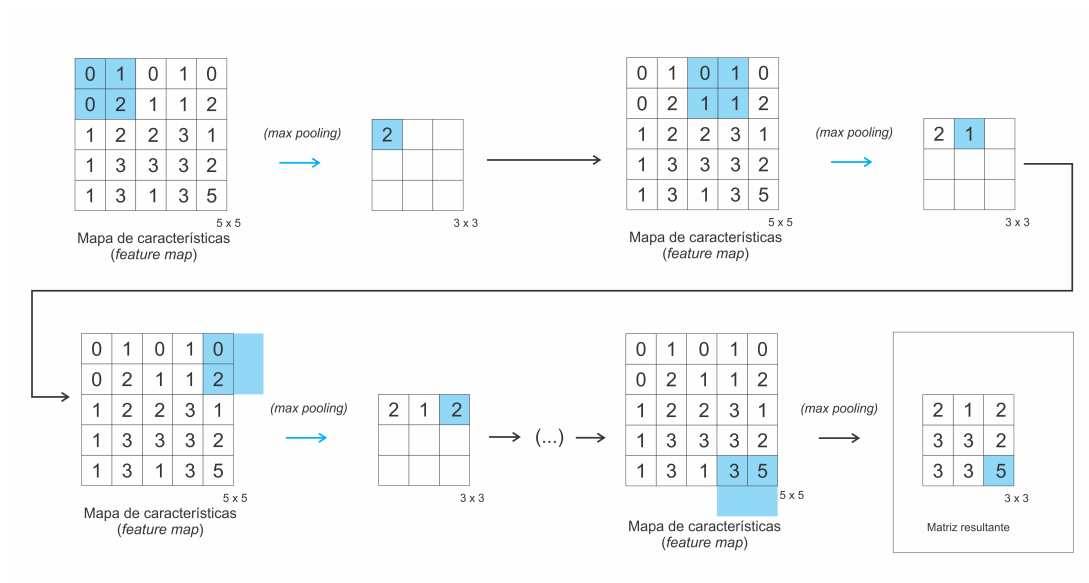
Figura 7 – Multiplicação das regiões da matriz da imagem de entrada pela matriz de kernel definida pela rede neural para gerar o mapa de características da camada de convolução.



FONTE: Elaborado pela autora.

O mapa gerado é utilizado como entrada na camada de *pooling*. Esta camada é responsável por encontrar características de alta relevância através da função *max pooling* que seleciona o maior valor de uma determinada região da matriz do mapa de características e o armazena em uma matriz, como demonstra a Figura 8. Como resultado, obtém-se uma matriz de tamanho consequentemente menor que porta as melhores características em sua composição (FARIA, 2018).

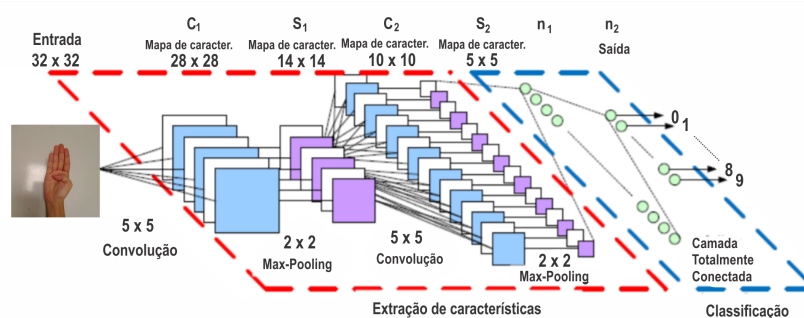
Figura 8 – Aplicação da função *max pooling* sobre o mapa de características produzido na etapa anterior (convolução) para gerar uma submatriz menor com características relevantes.



FONTE: Elaborado pela autora.

Por último, temos a camada totalmente conectada (*fully connected*) cujos neurônios da camada anterior estão conectados a todos os neurônios da camada seguinte (Figura 9 - Classificação). Ela usa as características obtidas nas camadas de convolução e *pooling* para fazer a classificação da imagem de entrada indicando a qual classe ela pertence com base no conjunto de treinamento (SILVA, 2018).

Figura 9 – Representação visual do processo de extração de características e classificação realizado pela CNN. Na extração de características (tracejado vermelho) há as etapas de convolução e *pooling* em alternância, enquanto na classificação (tracejado azul) há a camada totalmente conectada.



FONTE: Adaptado de (SILVA, 2018).

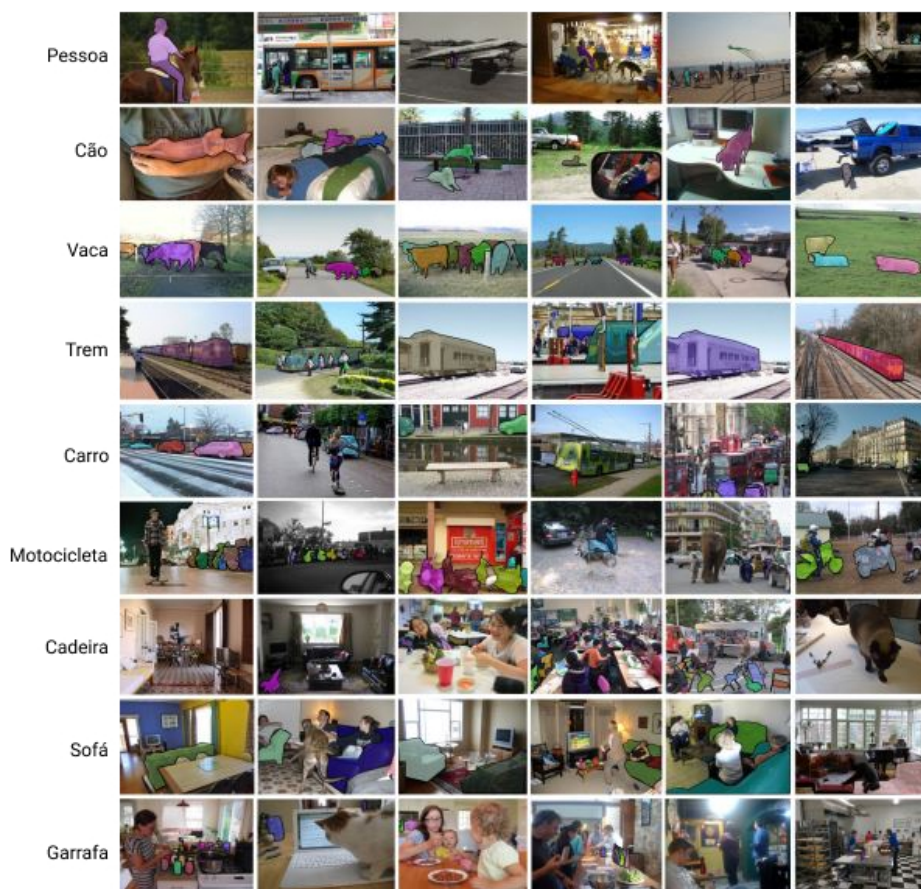
2.3.3 Microsoft Common Objects in Context - COCO

Uma rede necessita de uma base de dados para treinar seu modelo. Na implementação da plataforma descrita neste trabalho, foi utilizado um modelo pré-treinado com o *dataset Microsoft Common Objects in Context (COCO)*.

Trata-se de um conjunto de dados em grande escala amplamente utilizado. Tem mais de 330.000 imagens e 200.000 assuntos rotulados com pontos-chave como os exemplos da Figura 10. O conjunto de dados COCO não é apenas proposto para estimativa e análise de pose, mas também usado para detecção de objetos e segmentação de imagem em ambientes naturais e reconhecimento em contexto (ZHENG et al., 2021).

Ele foi desenvolvido para colaborar com o avanço do estado da arte em relação ao reconhecimento de objetos em cena considerando outros objetos que também fazem parte da imagem e não só o elemento central. Isso parte do pressuposto de que os sistemas atuais de reconhecimento funcionam bem para a visualização de objetos centrais, mas encontram dificuldades para reconhecer objetos parcialmente ocluídos ou em meio à desordem (cenas com vários objetos) (LIN et al., 2015).

Figura 10 – Amostras de categorias rotuladas do conjunto de dados do MS COCO



FONTE: Adaptado de (LIN et al., 2015).

2.3.4 Ferramentas web

A plataforma ArticuLi foi desenvolvida utilizando framework Django para desenvolvimento *backend* com a linguagem de programação Python, *frontend* com HTML, CSS e Bootstrap, e JavaScript para a reprodução de áudio na web através da *Web Speech API*.

O Django é um *framework* utilizado para o desenvolvimento de aplicações *web* que proporciona rapidez e facilidade no processo de implementação. Para isso, a ferramenta encapsula questões ligadas ao tratamento de requisições, mapeamento objeto-relacional, segurança e preparação de respostas HTTP (CARDOSO; BISPO et al., 2019). Além disso, o Django é caracterizado por ser versátil, pois funciona com outros frameworks e pode fornecer conteúdo em vários formatos (HTML, XML - *eXtensible Markup Language* - Linguagem de marcação extensível, JSON - *JavaScript Object Notation* - Notação de Objetos JavaScript, entre outros) (SOLIQUE; SASS; LIMA, 2018).

Como modelo de arquitetura o Django utiliza o padrão MTV (*Model-Template-View*). Nele, a camada *model* é responsável por fornecer o acesso aos dados e é onde são definidos os atributos e comportamentos das entidades que irão compor o sistema. Já a *views* é responsável pelo processamento das requisições dos usuários através de funções ou classes base conforme o roteamento de suas URLs. É na camada *template* que se define como os dados serão apresentados aos usuários, ou seja, é nela que a *interface* do sistema é construída (SOLIQUE; SASS; LIMA, 2018; CARDOSO; BISPO et al., 2019).

Para a *interface*, o Django é conhecido por gerar HTML dinamicamente que contém a parte estática comum, porém com algumas sintaxes especiais que indicam como o conteúdo dinâmico será inserido, por exemplo para variáveis que transmitem valores utiliza-se o marcador `{{variável}}` e para tags que controlam a lógica utiliza-se `{% tag %}` (CARDOSO; BISPO et al., 2019). Para compor a parte visual da *interface*, o CSS juntamente com o *framework* Bootstrap é utilizado para garantir um desenvolvimento *frontend web* de maneira mais rápida e com facilidade de adaptação à dispositivos de qualquer forma e tamanho (SOUZA; SANTO, 2020).

A linguagem Python escolhida é versátil e simples, ela permite o desenvolvimento ágil com redução de desperdícios, melhoramento na qualidade do código e otimização de performance. Trata-se de uma linguagem de alto nível altamente próxima do entendimento humano, modular necessitando de importações para uso, com vários componentes inclusos a sua biblioteca padrão e com coleções de dados bem elaboradas, como listas e tuplas e seus métodos auxiliares (SILVA; SILVA, 2019).

Já a *Web Speech API* fornece uma *interface JavaScript* para análise e síntese de fala a partir de texto e vice-versa para aplicativos da web. Ela possui arquitetura baseada em eventos, motivo que levou ao uso do JavaScript, e permite que os programas processem a fala de maneira assíncrona à execução de outras funcionalidades (ADORF, 2013).

A Figura 11, a seguir, mostra a *interface* do plataforma ArticuLi, produzido com base nas ferramentas apresentadas, rodando em servidor local em um dispositivo *mobile* para traduzir a palavra casa. Na primeira tela da imagem temos (a) página principal da plataforma, (b) tela de ação, (c) página principal da plataforma com fotografias no servidor e carregadas no campo da LIBRAS e (d) página de resultado da tradução e áudio.

Figura 11 – *Interface* responsiva da ArticuLi em tela de dispositivo *mobile* - tradução da palavra casa: (a) página principal da plataforma, (b) tela de ação, (c) página principal da plataforma com fotografias carregadas a partir do servidor e (d) página de resultado com tradução e opção de reprodução em áudio



FONTE: Elaborado pela autora.

3 TRABALHOS RELACIONADOS

Com o objetivo de entender e levantar dificuldades vivenciadas na construção de sistemas cujo foco está no estudo de classificadores para sinais da LIBRAS, este capítulo trará trabalhos atuais entre os anos 2017 e 2020.

De acordo com suas abordagens, os trabalhos com esse foco dividem-se em dois grupos: uso de visão computacional (processamento de imagens ou vídeos) e uso de luvas instrumentadas (utilização de sensores diversos) (DIAS et al., 2020). A saber, obedecendo a linha deste trabalho, focaremos na abordagem do primeiro grupo citado.

3.1 TRADUÇÃO DE LIBRAS POR IMAGEM

Utilizando conhecimentos de visão computacional, mais precisamente de processamento de imagem (PDI), (RIBEIRO; RODRIGUES, 2017) demonstram um sistema que identifica sinais do alfabeto manual da LIBRAS utilizando a biblioteca OpenCV e fazendo a captação das imagens via *webcam*.

Devido à escassez de bancos de imagens que retratam os sinais dessa língua, os autores montaram dois *datasets*, um com imagens positivas e outro com negativas. As positivas remetem ao sinal com veracidade, enquanto que as negativas servem para o sistema reafirmar o que não corresponde ao dado sinal possibilitando noções de diferenciação à aplicação. Cada sinal foi fotografado 10 vezes em profundidades diferentes totalizando 260 imagens positivas para uso de reconhecimento e aplicação do algoritmo de cascata de classificadores (abordagem de *machine learning*).

Como resultado, obteve-se a identificação de letras lidas e interpretadas a cada dez segundos (tempo destinado ao posicionamento do usuário) - as letras que estão fora desse intervalo foram desconsideradas. O sistema foi testado por três usuários onde utilizou-se a média geométrica dada por $x\sqrt{abc}$ para quantificar os dados obtidos (número de tentativas de cada letra e a quantidade de vezes que o sinal foi identificado). A média é diretamente proporcional ao conhecimento (bom, médio ou inexistente) do usuário apresentando valores de 5.0, 3.71 e 2.81 respectivamente, como resultados.

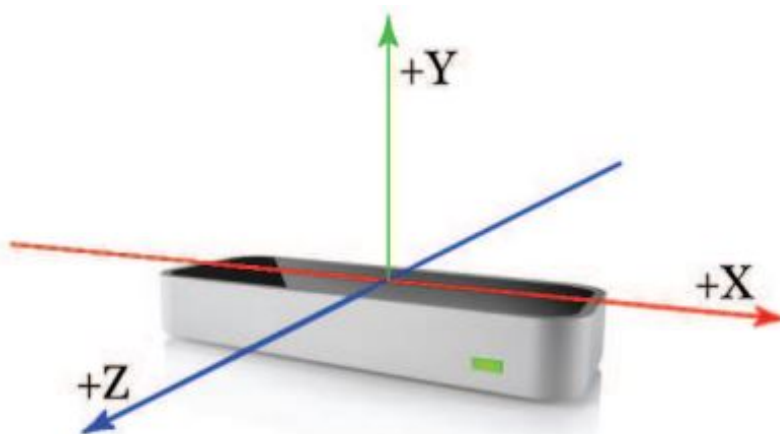
Ao observar as médias obtidas como resultado e sua relação com os níveis de conhecimento dos usuários, pode-se depreender que o sistema atende a sinais bem definidos, ou seja, executados por pessoas com conhecimento, mas apresenta dificuldades para identificar sinais com configuração inexata.

3.2 SINGAPP: UM MODELO DE IDENTIFICAÇÃO DE LÍNGUA DE SINAIS ATRAVÉS DE CAPTURA DE MOVIMENTO EM TEMPO REAL

Em seu trabalho, (LEAL, 2018) fala a respeito da inclusão de surdos e menciona o desconhecimento de ouvintes sobre a Língua Brasileira de Sinais ainda ser uma barreira. Nesse sentido, ferramentas computacionais trazem benefícios e terminam por ajudar na interação surdo-ouvinte.

Este trabalho configura-se, então, em um modelo de reconhecimento de sinais provenientes da captura da estrutura da mão através de um aparelho específico: o *Leap Motion Controller* (LMC). A Figura 12 retrata o aparelho onde ele possui sensor infravermelho e câmeras que fazem o reconhecimento e rastreamento dos dedos das mãos de forma simultânea dentro da área de captura que ele alcança. Para o estudo, foram escolhidos os parâmetros da configuração de mão e movimento.

Figura 12 – Leap Motion Controller e coordenadas cartesianas X, Y e Z utilizadas no estudo.



FONTE: Adaptado de (LEAL, 2018).

Para a classificação da configuração de mão, utilizou-se uma *Multilayer Perceptron Neural Network* que tinha como entradas as distâncias euclidianas entre a posição de cada dedo e a posição do centro da palma, além das distâncias euclidianas entre a posição da ponta de dois dedos consecutivos.

Em relação aos movimentos, uma segunda rede neural foi utilizada. Os dados de entradas passados correspondem ao cálculo de distância euclidiana do centro da palma da mão capturada a cada 30 frames e o cálculo do ângulo do percurso do centro da palma a cada 60 frames.

Com base de dados própria (114.000 amostras das 19 configurações de mão de 3 pessoas diferentes), os testes realizados utilizaram de 1 a 50 neurônios tendo melhor aproveitamento quando com 15 neurônios atingindo 99,8% de acerto. Entretanto, o aparelho não conseguiu identificar de maneira satisfatória a configuração de mão utilizada para as letras H, K e P.

O teste para reconhecimento de movimentos contou com 300 amostras de cada um

dos 6 movimentos classificados, cada amostra com 9 informações diferentes. Como resultado, obteve-se 86,7% de precisão.

Nesse contexto, destaca-se a necessidade da presença de um dispositivo não usual como o LMC implicando gastos extras para a aplicação do trabalho e melhorias para que a identificação das letras H, K e P possa acontecer.

3.3 CLASSIFICAÇÃO AUTOMÁTICA DE SINAIS VISUAIS DA LÍNGUA BRASILEIRA DE SINAIS REPRESENTADOS POR CARACTERIZAÇÃO ESPAÇO-TEMPORAL

Segundo (MACHADO et al., 2018), pessoas portadoras de deficiência auditiva possuem diversos problemas de acessos aos serviços de educação e saúde, aos quais tem direito. Há variáveis que influenciam a compreensão entre as pessoas que se comunicam, assim como há a necessidade da existência de intérpretes no momento de interação.

Essa dissertação de mestrado teve como objetivo a classificação de vídeos da Libras, através da utilização de características espaço-temporais aprendidas por uma *Neural Network*, com uso de transferência de aprendizado. Ela abrange o método da visão computacional considerando a produção de soluções mais baratas e não intrusivas.

O treinamento da *Deep Neural Network* deste trabalho deu-se através da base de dados ISOGD. Ela recebe como entrada vídeos e, a partir disso, faz a extração de quadros RGB, com profundidade e fluxo óptico adequados. Ao longo do estudo, o modelo original da rede foi alterado.

Para a geração da base de dados foram selecionados 7 intérpretes voluntários (3 homens sendo 1 surdo e 4 mulheres sendo duas surdas) onde cada profissional repetiria 6 vezes cada sinal selecionado, com isso seria totalizado 21.420 vídeos. Todos os vídeos filmados correspondem a cerca de 35 horas de duração. Para armazenamento, conversão dos vídeos em quadros de RGB e de profundidade foram utilizados 2 discos rígidos adicionais de 3 TB.

Devido a inconsistências no momento de rotulagem e gestos não correspondentes a sinais da Libras feitos pelos intérpretes durante as gravações, as amostras utilizadas nessa pesquisa contemplaram cerca de 84 sinais, 6 repetições cada, feitos por 7 intérpretes diferentes. Ou seja, a amostra utilizada possuía 3.528 vídeos.

Os resultados obtidos durante o treino da base de amostra da Libras, com uso de transferência de aprendizado apresentaram acurácia de 58,65%, 69,04% e 79,80% para as bases ISOGD (modelo original), Libras (transferência de aprendizado/10 épocas) e Libras (transferência de aprendizado/20 épocas) respectivamente. Nesse sentido, constatou-se a necessidade do refinamento das épocas para aumento de significância nos resultados por vídeo.

No trabalho, observa-se esforços em relação a rotulagem dos frames dos vídeos onde os sinais encontram-se em execução, isso abrange atividade exaustiva com foco em repetição.

3.4 SISTEMA EMBARCADO LINUX PARA RECONHECIMENTO DE VOGAIS EM LIBRAS UTILIZANDO APRENDIZAGEM DE MÁQUINA

De modo a promover a acessibilidade e entendendo que as tecnologias da comunicação facilitaram a comunicação humana e as formas de interação, (CAVALCANTE et al., 2019) propõem facilitar a tradução simultânea da Libras para o português brasileiro através do reconhecimento automático dos gestos de vogais utilizando o Sistema Embarcado Linux instalado em um *Raspberry Pi* e integração com o *Telegram*. Para isso, os sinais foram capturados por uma câmera e identificados por um modelo de *machine learning* para detecção automática de objetos.

Em sua pesquisa documental não foram encontrados trabalhos envolvendo a aplicação de sistemas embarcados no dado contexto. Por isso, o levantamento foi direcionado a busca por trabalhos que utilizam sistemas embarcados em tarefas de processamento e reconhecimento de imagens, todos com *Raspberry Pi*.

As etapas realizadas para o desenvolvimento da proposta foram: criação de um banco de imagens, organização das caixas limitadoras para indicar onde os símbolos deveriam ser posicionados nas imagens, treinamento do modelo de aprendizagem profunda com *Single Shot Multibox Detector - SSD*, configuração do *Raspberry* (classificação das imagens, identificação dos sinais e envio das mensagens ao *Telegram*).

A integração com o *Telegram* se deu através de um *bot*, canal que possibilitou o envio das vogais reconhecidas pelo sistema. Os sinais traduzidos eram concatenados e, portanto, formavam uma *string*. Esta era enviada após a solução reconhecer o sinal de envio (definido pelos autores).

Nos experimentos computacionais foram feitos testes com imagens separadas para validar se os modelos estavam devidamente treinados. Nisso, verificou-se taxa de acerto de 96,67%. Entretanto, os testes possuem variâncias devido a dificuldades de reconhecimento, influência de luminosidade e distância, além da variação de quadros por segundos da Câmera Pi entre 11 e 17, fazendo com que muitos quadros desfocados fossem passados pelo algoritmo.

Concluiu-se que o reconhecimento em tempo real não é completamente assertivo podendo ser melhorado com uma base robusta de dados. Apresentou-se como ponto positivo o uso de *Raspberry Pi* ressaltando seu alto poder de processamento para diversas soluções e capacidade de ser aplicado na resolução de problemas variados.

O trabalho se limita ao reconhecimento de vogais e utiliza equipamento auxiliar diferente do dispositivo móvel do usuário para ser executado.

3.5 UM MÓDULO DE SOFTWARE PARA O RECONHECIMENTO DE SINAIS ESTÁTICOS DA LÍNGUA BRASILEIRA DE SINAIS

Ressaltando a comunicação como fator crucial para que um indivíduo esteja incluso na sociedade, (SANTOS et al., 2020) discorrem sobre a existência de aplicações que auxiliam e

facilitam a tradução da Língua Portuguesa para a LIBRAS através de avatar 3D. Entretanto, chamam a atenção para a escassez de ferramentas que façam o processo contrário, tradução da LIBRAS para o Português.

Geralmente, as abordagens utilizadas para o reconhecimento de sinais por computador compreendem uma luva com dispositivos (sensores) agregados para detecção e rastreamento de movimentos, além de técnicas de visão computacional. Estas configuram-se como métodos não invasivos por não exigirem acessórios acoplados ao corpo ou gastos adicionais. Optou-se, então, pela segunda abordagem.

Trata-se do desenvolvimento de um módulo de *software* capaz de traduzir sinais estáticos da LIBRAS para a Língua Portuguesa. Para tanto, utilizou-se o algoritmo SURF para detecção de características invariantes, rotação e escala (extração de características da imagem), câmera RGB comum para captura de fotografias, treinamento de *neural network* com WEKA utilizando *Backpropagation* (aprendizagem supervisionada) e OpenCV para o processamento de imagem.

Os testes foram realizados com uma amostra composta por duas mil (2000) imagens de tamanho 200x200 pixels em escala de cinza. Cada imagem representa a configuração de mão dos números de zero a nove na LIBRAS, sendo utilizada para cada configuração de mão uma base de dados de 200 imagens. As imagens foram obtidas de duas pessoas diferentes e com pequenas rotações de gestos variando entre mão direita e esquerda.

Na execução do treinamento da *neural network* utilizou-se a técnica de *cross-validation* com o método padrão de 10-*folds*. No final, a média de reconhecimento encontrado no modelo foi de 80,45%, sendo 1609 das 2000 imagens reconhecidas de forma correta e 391 incorretas (19,55%).

Com isso, o estudo apresentado limita-se à identificação de sinais estáticos que representam números e não abrangem sinais que retratam as letras do alfabeto manual.

4 ARTICULI - PROPOSTA DE PLATAFORMA DE TRADUÇÃO DE DATILOLOGIA COM BASE EM VISÃO COMPUTACIONAL PARA FORMAÇÃO DE PALAVRAS

A Articuli consiste em uma plataforma *web* que recebe um conjunto de fotos da mão do usuário em sequência com configurações que representam as letras do alfabeto datilológico para, então, realizar as traduções e concatená-las a fim de retornar a palavra desejada na língua portuguesa.

O nome da plataforma apresentado na Figura 13 surgiu da relação entre a abordagem usada na aplicação (mapeamento dos pontos-chave da mão para estimativa da sua pose) que é representada na logo pelas letras I ligadas, o verbo articular, com seu significado de produzir os sons da fala ou articular uma palavra, e a sigla LIBRAS, onde são usadas as duas primeiras letras em alusão a Língua Brasileira de Sinais. Como resultado, obteve-se o nome Articuli que faz referência ao verbo articule da Língua Portuguesa cuja conjugação está no imperativo afirmativo e transmite a ideia de estímulo a ação para o público portador de deficiência auditiva ao conhecerem a plataforma apresentada.

Figura 13 – Logo da plataforma Articuli



FONTE: Elaborado pela autora.

A seguir, são apresentados todos os processos que colaboraram para a construção da aplicação considerando etapas que vão desde a modelagem da plataforma à implementação partindo da visão anatômica da mão humana, mapeamento de *key points* com OpenCV, análise e definição de características, formação de palavras até as ferramentas *web* utilizadas.

4.1 MODELAGEM DA PLATAFORMA

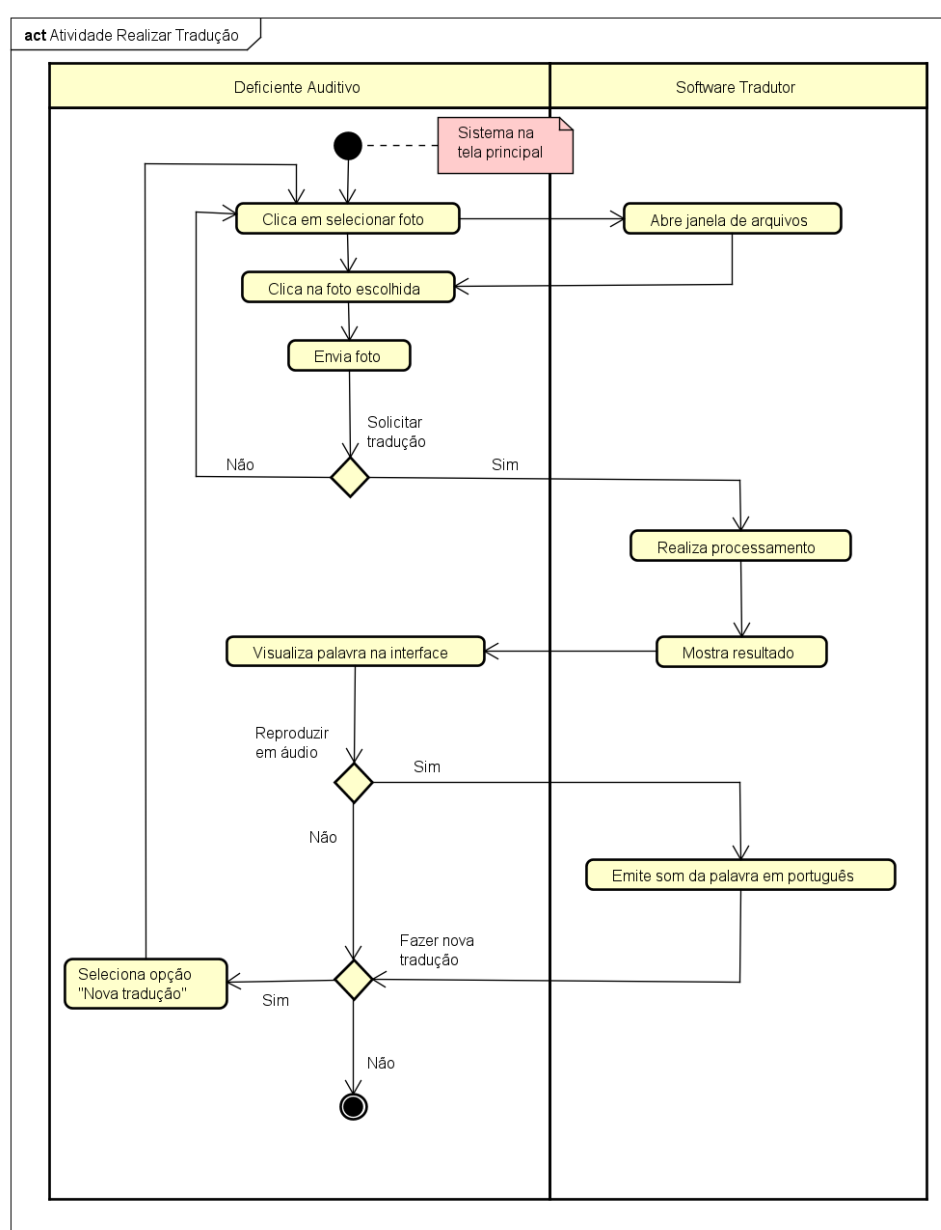
Os modelos conceituais partem da abstração de aspectos reais e compõem-se de estruturas que remetem à implementação. Geralmente esses modelos são representados utilizando a *Unified Modeling Language* (UML) (SILVA; FARINELLI; ALMEIDA, 2017). Como artefato dessa linguagem, o Diagrama de Atividade foi escolhido para representar as ações que compõem o fluxo principal da aplicação e quem às realiza.

Na sequência, o *Service Blueprint* ou Plano de Serviço, dentre suas diversas funções como artefato, foi utilizado para a visualização geral da arquitetura do sistema. Ele demonstra os pontos de contato do deficiente auditivo em sua jornada de interação com a interface e a execução paralela dos processos internos invisíveis e necessários para a entrega da palavra traduzida.

4.1.1 Diagrama de Atividade

A Figura 14 demonstra o fluxo principal juntamente com os atores e cada ação que executam onde, na tela principal do software, o deficiente auditivo escolhe fotos em sequência e às envia para a aplicação. Em seguida, ele solicita a plataforma a tradução dos sinais expressos nas fotos. Esta, por sua vez, realiza o processo de tradução e retorna o resultado em português na forma de palavra. A partir daí, tendo visualizado o resultado, o usuário poderá optar por reproduzi-lo sonoramente ou não e iniciar uma nova tradução, caso queira.

Figura 14 – Diagrama de Atividade representando as ações do fluxo principal Realizar Tradução e os autores que o executam



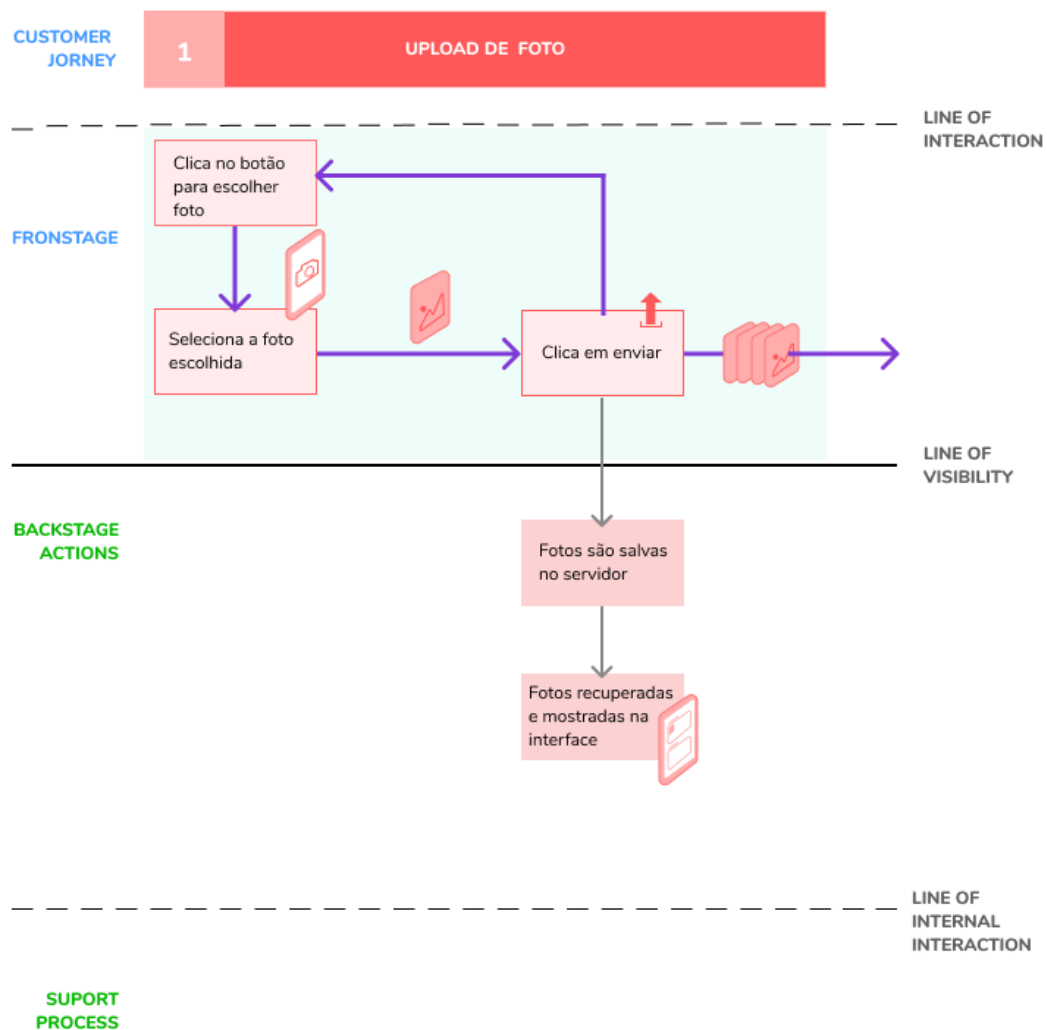
FONTE: Elaborado pela autora.

4.1.2 Service Blueprint

A seguir, as Figuras 15, 16, 17 demonstram em detalhes como funciona a logística que o sistema deve atender em sua implementação com camadas internas explícitas. O *Service Blueprint* mostra cada etapa da jornada (*customer journey*), os pontos de contato entre usuário e interface (*fronstage*), as ações internas ou de bastidores (*backstage actions*) após a linha de visibilidade (*line of visibility*) e os processos fundamentais para que as ações internas possam ser executadas com sucesso (*support process*), caso sejam necessários.

A Figura 15 mostra a primeira etapa de interação do usuário com a plataforma. Nesse momento, ele realiza o *upload* de fotos diretamente do seu celular. Ao realizar os envios, as fotos ficam salvas no servidor e são retornadas para a interface.

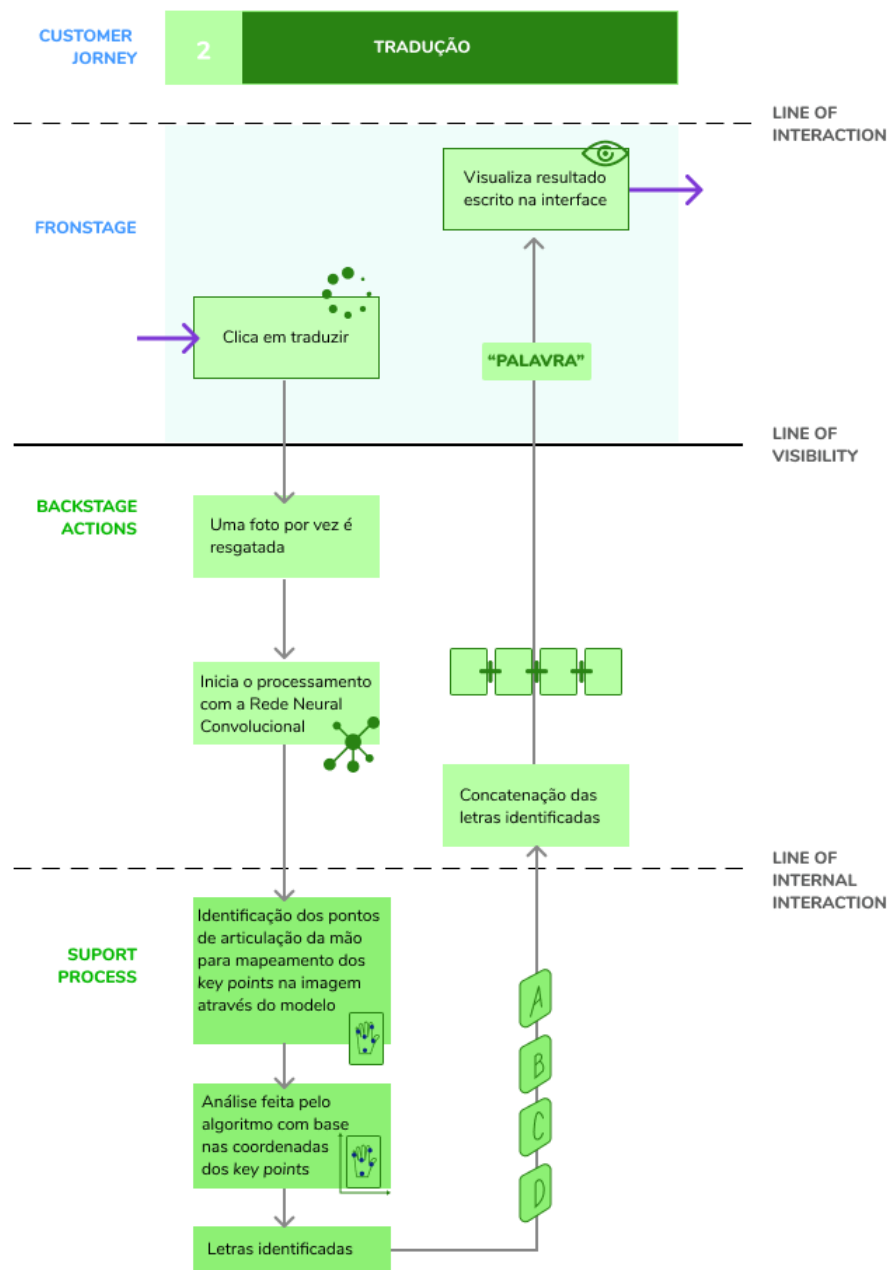
Figura 15 – Primeira etapa da interação do usuário em sua jornada - Upload de foto



FONTE: Elaborado pela autora.

Já a Figura 16 mostra a segunda etapa responsável pela tradução. Ao solicitar a tradução, o usuário aguarda o processamento. No *backstage* da aplicação uma foto por vez é resgatada e submetida ao processamento com a Rede Neural Convolucional. Mais a fundo, no *support process*, as coordenadas dos pontos de articulação são mapeadas, em seguida analisadas e comparadas para identificação das respectivas letras. Em seguida, as letras identificadas são concatenadas para formar a palavra em português que, por sua vez, será o resultado mostrado na interface.

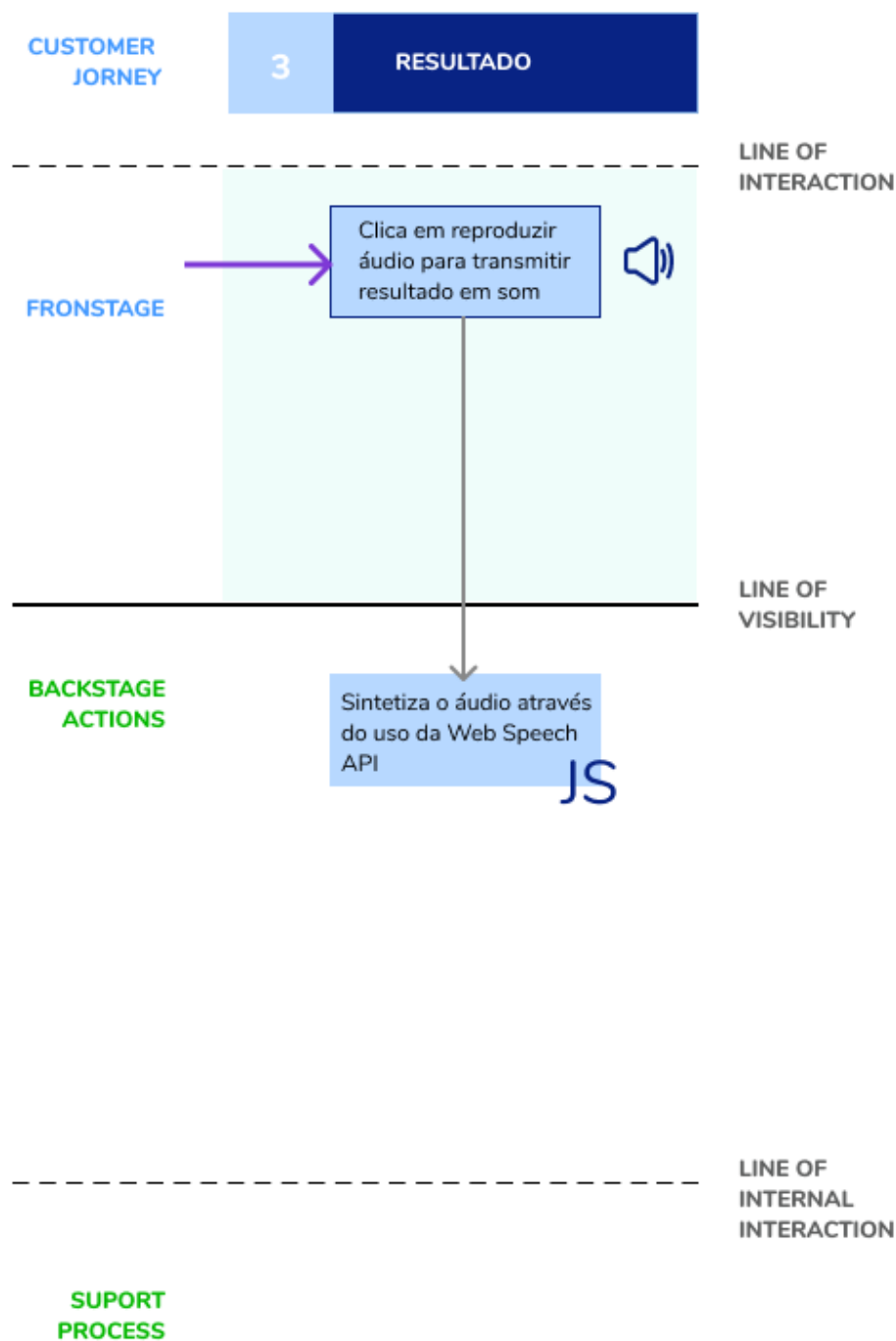
Figura 16 – Segunda etapa da interação do usuário em sua jornada - Tradução



FONTE: Elaborado pela autora.

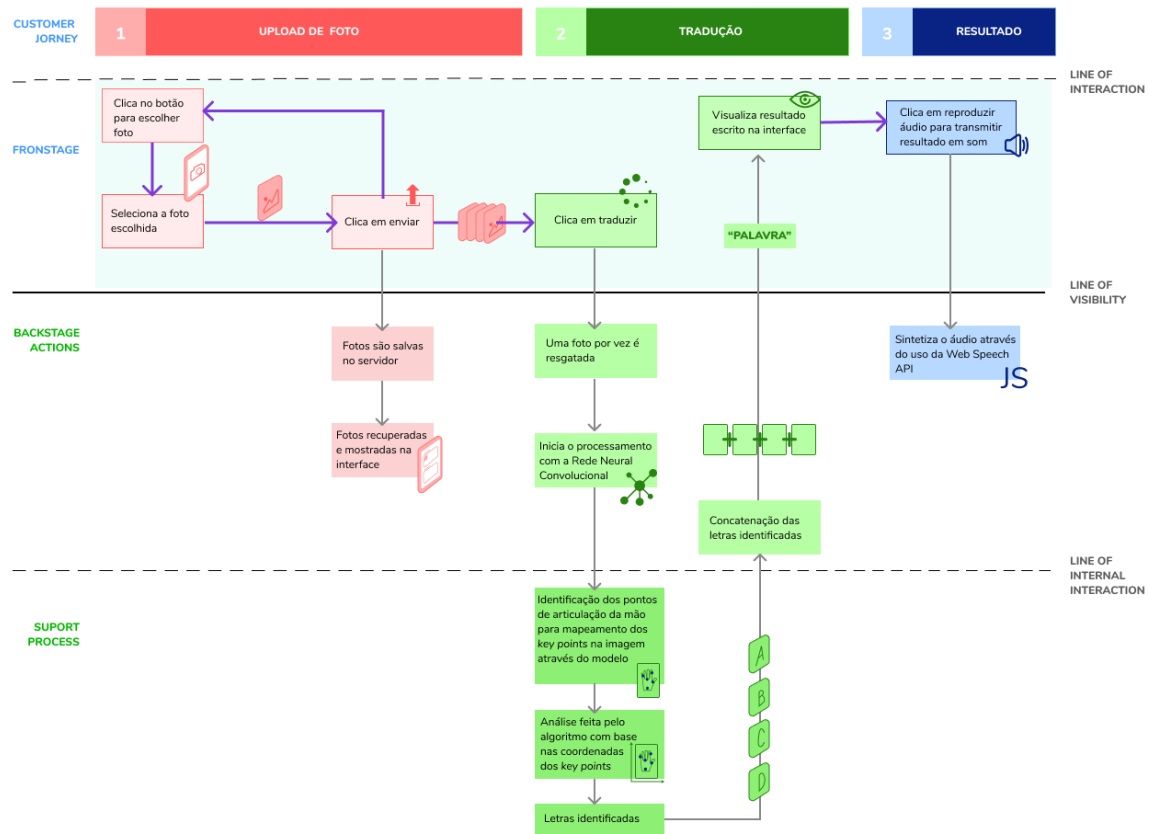
A Figura 17 demonstra o momento final da interação, o resultado. O usuário deficiente auditivo, ao clicar em reproduzir, faz com que a plataforma emita o som referente a palavra sintetizado através do serviço da *Web Speech API*.

Figura 17 – Terceira etapa da interação do usuário em sua jornada - Resultado



Por último, pode-se observar na Figura 18 todas as etapas mostradas anteriormente em sequência: o *upload* de foto, a realização da tradução e o resultado com opção de reprodução em áudio.

Figura 18 – *Service Blueprint* completo - etapas de *upload* da foto, tradução e resultado em sequência.



FONTE: Elaborado pela autora.

4.2 IMPLEMENTAÇÃO

No processo de desenvolvimento da plataforma, utilizou-se conhecimentos sobre Visão Computacional, desenvolvimento *web* responsivo, *framework* Django, linguagem de programação Python, HTML, CSS e JavaScript. Para a construção do processamento central da plataforma utilizou-se um modelo pré-treinado de Rede Neural Convolutiva com base no conjunto de dados da *Microsoft Common Objects in Context* e uso da biblioteca OpenCV para a localização dos *key points* da mão na fotografia.

A seguir, o processo de implementação é descrito partindo do conhecimento sobre a visão anatômica da mão humana, mapeamento de *key points* com OpenCV, análise e definição de características, formação de palavras e ferramentas *web* utilizadas.

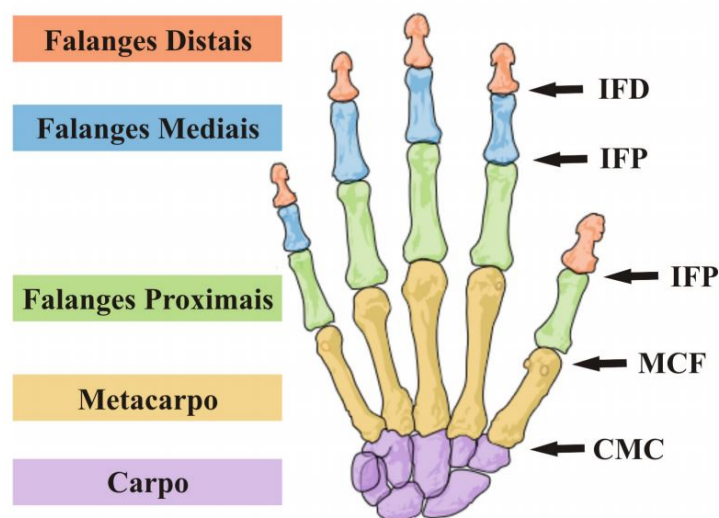
4.2.1 Visão anatômica da mão humana

Além dos conhecimentos técnicos da área de tecnologia e explicações sobre a LIBRAS, levou-se em consideração conhecimentos voltados para a anatomia da mão humana uma vez que a aplicação busca identificar os pontos de articulação desse membro, durante a análise das imagens.

A estrutura da mão é composta por 27 ossos classificados anatomicamente em três grupos distintos: carpo, metacarpo e falanges. O carpo é composto por oito ossos que constituem a base da mão (pulso), dentre estes atendem à linha proximal o escafoide, semilunar, piramidal e pisiforme e à linha distal o trapézio, trapezoide, capitato e hamato. O metacarpo é composto por cinco ossos que são numerados a partir do polegar e formam a palma da mão. Já as falanges abrangem a região dos dedos e subdividem-se em proximal, distal e medial, onde apenas o polegar não possui a falange medial (OLIVEIRA et al., 2016; MALHEIROS, 2019).

A seguir, a Figura 19 mostra que, conforme sua localização, a articulação pode ser nomeada em Carpometacarpal (CMC) - entre os ossos da linha distal e os ossos do metacarpo, Metacarpofalângica (MCF) - entre os ossos do metacarpo e as falanges proximais, Interfalângica Proximal (IFP) - entre as falanges proximais e as falanges mediais e Interfalângica Distal (IFD) - entre as falanges mediais e distais (OLIVEIRA et al., 2016).

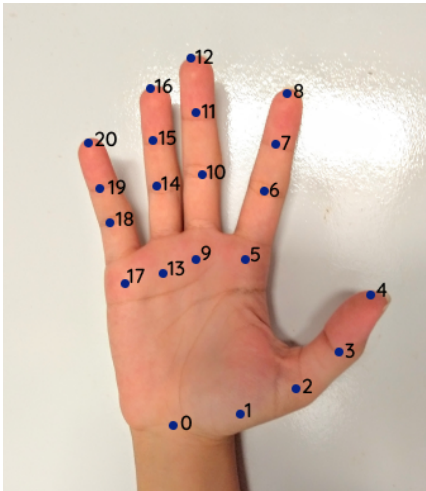
Figura 19 – Anatomia da mão - localização das articulações CMC, MCF, IFP e IFD.



FONTE: (OLIVEIRA et al., 2016).

Já a Figura 20 mostra os pontos mapeados que representam a maioria das articulações usadas durante a estimativa da pose da mão pelo modelo utilizado na aplicação. Entre as articulações Carpometacarpais, apenas a do polegar é mapeada enquanto às dos demais dedos (indicador, médio, anelar e mínimo) seguem direto do ponto 0 - base da mão (carpo ou pulso) até as suas articulações Metacarpofalângicas (pontos 5, 9, 13, 17 respectivamente).

Figura 20 – Representação dos pontos de articulações da mão que são mapeados pelo modelo da COCO na abordagem da estimativa da pose da mão.



FONTE: Elaborado pela autora

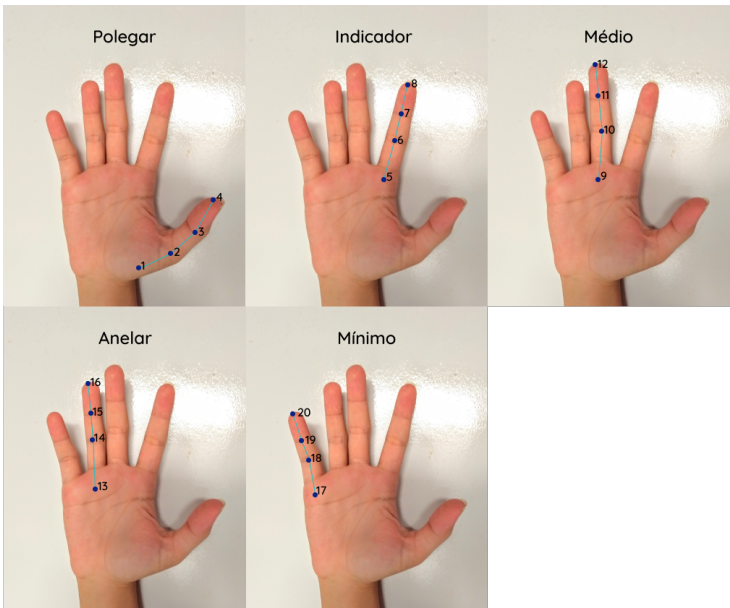
Os dedos são, então, identificados com base nos intervalos sem contar com o pulso (ponto 0). A Tabela 2 descreve cada intervalo e a Figura 21 os demonstra visualmente.

Tabela 2 – Intervalos de identificação de cada dedo da mão

Dedo	Intervalos
Polegar	1 a 4
Indicador	5 a 8
Médio	9 a 12
Anelar	13 a 16
Mínimo	17 a 20

FONTE: Elaborada pela autora

Figura 21 – Representação dos intervalos dos dedos da mão



FONTE: Elaborado pela autora

4.2.2 Mapeamento de *key points* com OpenCV

A OpenCV realiza a leitura do modelo da CNN pré-treinada através do arquivo .prototxt, que possui a descrição da arquitetura da CNN, e do arquivo .caffeModel que possui a CNN reconhecida. Em seguida, recebe imagens e as converte para um formato aceito na entrada da rede. Com isso, os dados são passados de camada a camada até que as saídas sejam produzidas.

O modelo utilizado conta com 22 *key points*. A mão contém 21 (do 0 ao 20), como demonstrado na subseção anterior com a Figura 20, acrescido do fundo da imagem como o ponto de número 22. Conforme mostrado na Figura 22, os *key points* encontrados são representados através de coordenadas com os valores dos *pixels* do eixo x e eixo y da imagem e armazenados em formato de tupla.

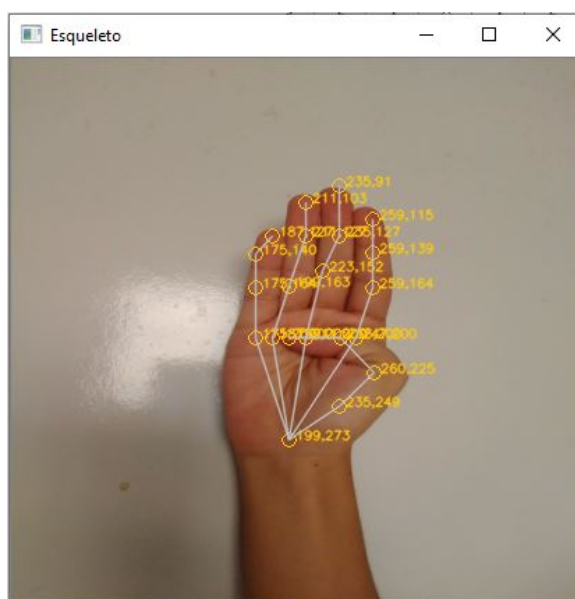
Figura 22 – Mapeamento de *key points* da mão em configuração do sinal B



FONTE: Elaborado pela autora.

Após o mapeamento, as coordenadas dos *key points* são interligados com linhas formando o esqueleto da configuração que foi projetada, conforme pode-se observar na Figura 23 que traz o exemplo do sinal da letra B.

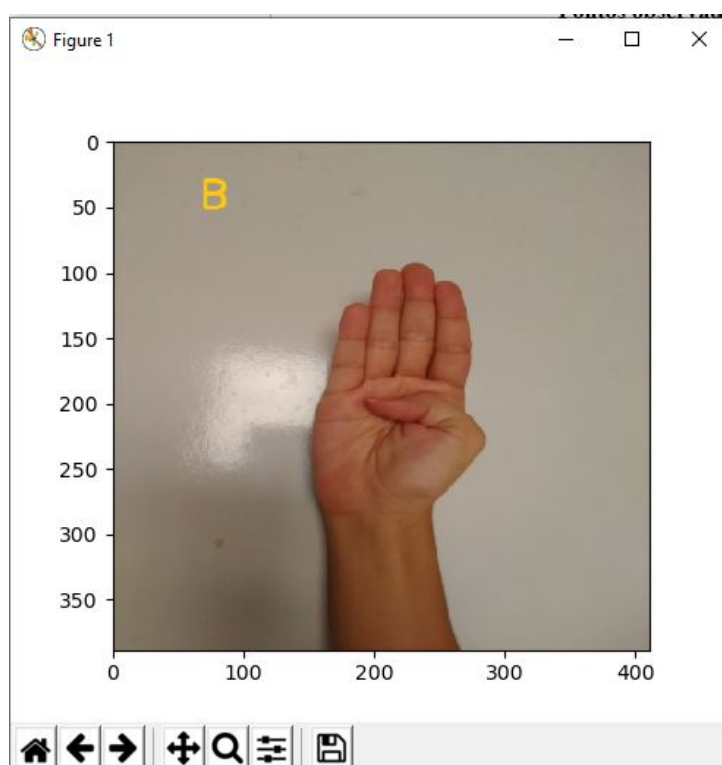
Figura 23 – Esqueleto formado com a ligação das coordenadas de cada *key point* da mão executando sinal de B



FONTE: Elaborado pela autora.

Em seguida, o algoritmo realiza análises que serão explicadas na próxima subseção envolvendo as alturas, posições e proximidades entre os pontos retornando características que, juntas, identificam a letra correspondente ao sinal da datilologia. A Figura 24 mostra o resultado final continuando o exemplo da letra B.

Figura 24 – Resultado do processo de identificação do sinal da letra B



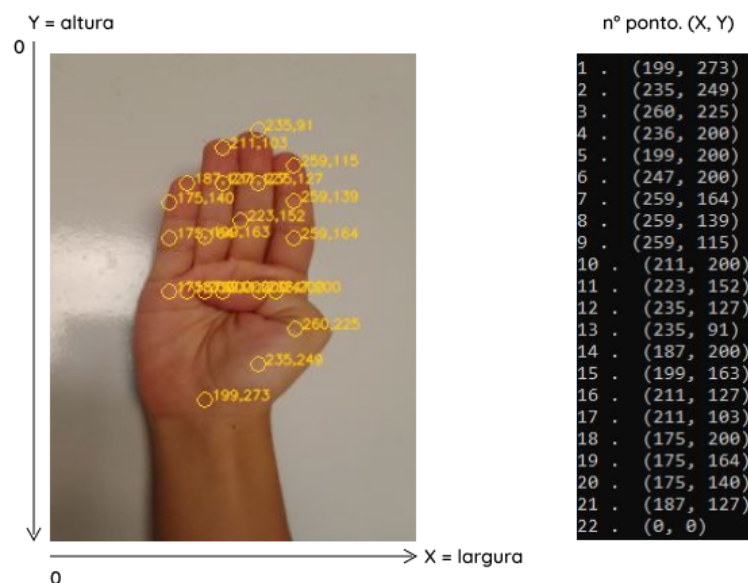
FONTE: Elaborado pela autora.

4.2.3 Análise e definição de características

Para chegar ao resultado, o algoritmo conta com três módulos base essenciais que retornam informações para a identificação do sinal em português, são eles: módulo altura, proximidade e posição.

O algoritmo inicia com o módulo altura que recebe o conjunto das coordenadas x e y para localizar aquelas que correspondem às pontas de cada dedo, ou seja, que correspondem aos pontos 4, 8, 12, 15 e 20 guardando o valor de y. No plano cartesiano, esse valor corresponde a altura e será comparado com o valor y do ponto 0 (pulso). A Figura 25 retrata a imagem no plano cartesiano e as coordenadas dos pontos encontradas no exemplo. É importante lembrar que em visão computacional quanto menor o valor de y, mais alto o ponto está na imagem e vice-versa. Portanto, o módulo retorna “acima” se o valor da ponta dos dedos for menor que o da altura do pulso e “abaixo” se for maior.

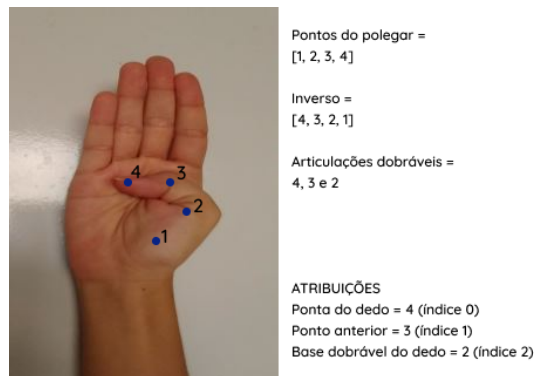
Figura 25 – À esquerda um exemplo do plano cartesiano com o sinal da letra B e pontos mapeados. À direita estão as coordenadas X e Y desses pontos.



FONTE: Elaborado pela autora.

Em seguida, o módulo posição recebe o intervalo de pontos que define cada dedo, sua denominação (polegar, indicador, médio, anelar ou mínimo) e o retorno do módulo altura, ou seja, se a mão está voltada para cima ou para baixo. Esse módulo verifica se os dedos estão dobrados, esticados horizontalmente ou esticados verticalmente. Ele leva em consideração as articulações que possuem qualidade dobrável (apenas 3 dos 4 pontos de cada intervalo dos dedos), e atribui um valor de base dobrável, valor de ponto anterior e valor de ponta do dedo resgatados do intervalo invertido. Para comparação, o valor de ponto anterior é tratado como parâmetro e o valor de base e ponta do dedo é usado para encontrar os valores determinantes. A Figura 26 demonstra a explicação de forma objetiva.

Figura 26 – Exemplo prático dos dados usados no módulo posição - Dedo polegar



FONTE: Elaborado pela autora.

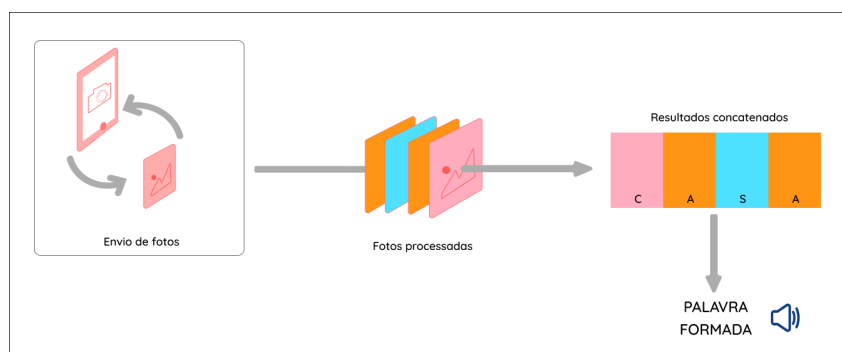
Já o módulo proximidade verifica os valores das coordenadas x de cada dedo comparando-os. Seu objetivo é identificar se os dedos estão próximos ou afastados e se estão na mesma posição (esticados ou dobrados) utilizando sempre um dedo posterior como parâmetro de comparação, com exceção do mínimo, que compara-se ao anelar (dedo anterior). Assim, o polegar compara-se ao indicador, o indicador ao médio, o médio ao anelar, o anelar ao mínimo e o mínimo ao anelar, nessa ordem.

O conjunto das características retornadas descreve como os dedos se portam quando um sinal é executado, desse modo é possível associá-lo a uma letra. O algoritmo contém todas as letras do alfabeto caracterizadas, portanto, para finalizar, ele relaciona as informações e identifica que letra está sendo expressa na imagem.

4.2.4 Formação da palavra

Após o processamento, cada letra resultante (identificada) é concatenada à posterior seguindo a sequência lógica das fotos de entrada. A concatenação, de forma mais clara, é o momento em que a palavra é formada. Portanto, após a finalização do processo, a palavra é apresentada ao usuário. A Figura 27 mostra exclusivamente o fluxo das fotos para a formação das palavras.

Figura 27 – Fluxo das fotos na plataforma ArticuLi



FONTE: Elaborado pela autora.

5 RESULTADOS E DISCUSSÕES

Para verificar a eficiência da plataforma na tarefa de identificação dos sinais da LIBRAS em fotografias e visualizar seus pontos de melhoria, esse capítulo traz resultados de testes efetuados com base na metodologia apresentada.

5.1 CONSTRUÇÃO DO BANCO DE IMAGENS

Para execução dos testes foi utilizado um conjunto de imagens fotografadas por 10 pessoas, onde apenas 3 possuíam conhecimento em LIBRAS. Para montagem do conjunto, foi solicitado aos voluntários que fotografassem os sinais do alfabeto manual usando a câmera do seu próprio *smartphone* e efetuassem o envio dos registros por e-mail.

Os voluntários foram guiados através de um pdf elaborado pela autora que os instruiu a fotografar os sinais seguindo 3 critérios: não fotografar o sinal com sombra próxima aos dedos, deixar pulso aparente e posicionar a mão de forma a permitir a visualização das articulações. Com isso, obteve-se um conjunto balanceado com 5 amostras para cada sinal totalizando 130 imagens. Elas estão armazenadas em um repositório no github da autora, acessível no endereço eletrônico <https://github.com/andressamesquita/ImagensLibrasParaTeste>.

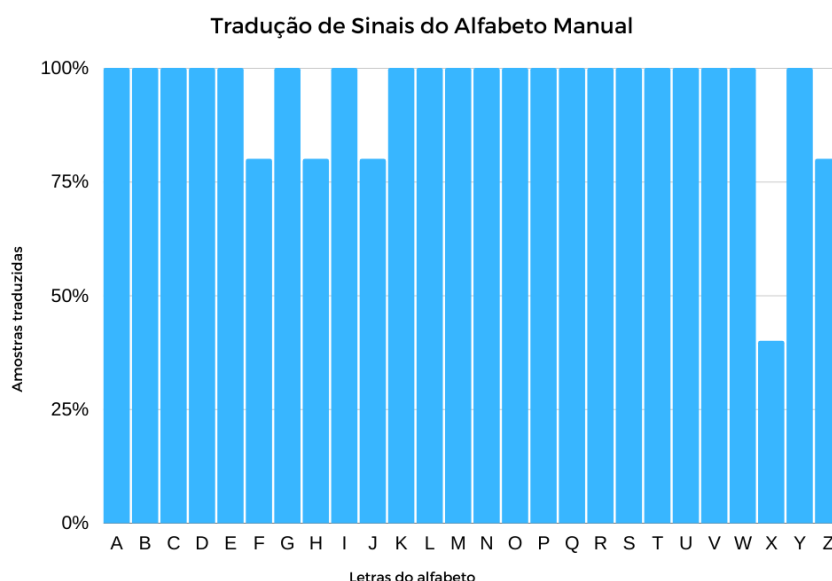
5.2 TRADUÇÃO DOS SINAIS

A plataforma do trabalho abrange a identificação dos sinais do alfabeto manual independente da dinâmica, seja ela existente ou não, ou seja, ela abrange a tradução de sinais tanto estáticos quanto dinâmicos trazendo alguns desafios com a abordagem de estimativa da pose da mão.

Considerando o número total de imagens e a quantidade de sinais traduzidos corretamente, incluindo os sinais estáticos e dinâmicos, alcançou-se uma acurácia de 94,6%, ou seja, das 130 imagens analisadas pela ferramenta, 123 foram traduzidas corretamente.

Agora, considerando a tradução de cada letra, a Figura 28 apresenta a ilustração de um gráfico que traz os resultados alcançados com a plataforma usando as imagens da base de dados. Na vertical, em azul, estão as barras que representam a quantidade de amostras traduzidas com sucesso sendo o máximo 5, como citado anteriormente, pois é o número de amostras por sinal. Na horizontal, cada barra corresponde a uma letra do alfabeto.

Figura 28 – Ilustração de gráfico com a representação da quantidade de amostras de sinais traduzidos corretamente pela ArticuLi.

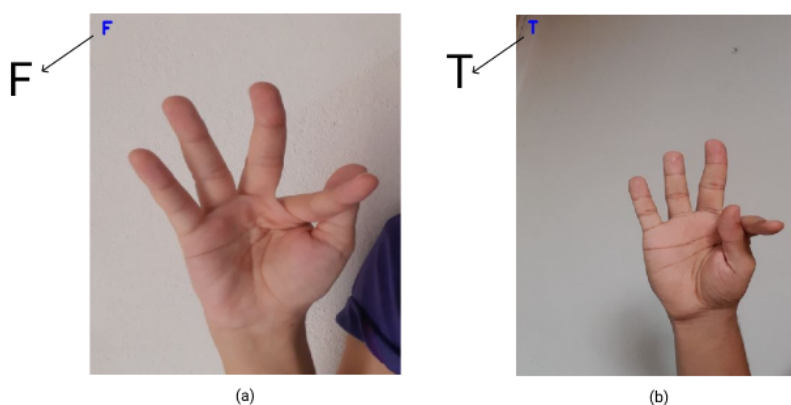


FONTE: Elaborado pela autora.

Como observado na Figura 28, as letras A, B, C, D, E, G, I, K, L, M, N, O, P, Q, R, S, T, U, V, W e Y obtiveram todas as amostras traduzidas atingindo 100% de acurácia. Já as letras F, H, J e Z chegaram a 80% e X a 40% de acurácia.

Durante os testes, o sinal de F, com 80% de acurácia, não obteve sucesso em uma das amostras. Naquela não identificada, a configuração da mão apresentada na fotografia confundia-se com a configuração do sinal da letra T. Como observado na Figura 29, o sinal dessas letras possuem configuração da mão semelhante e são diferenciados pela posição do polegar. Se este estiver antes do dedo indicador quer dizer que aquele sinal refere-se a letra T, já se estiver posicionado após o dedo indicador ele se refere ao sinal da letra F.

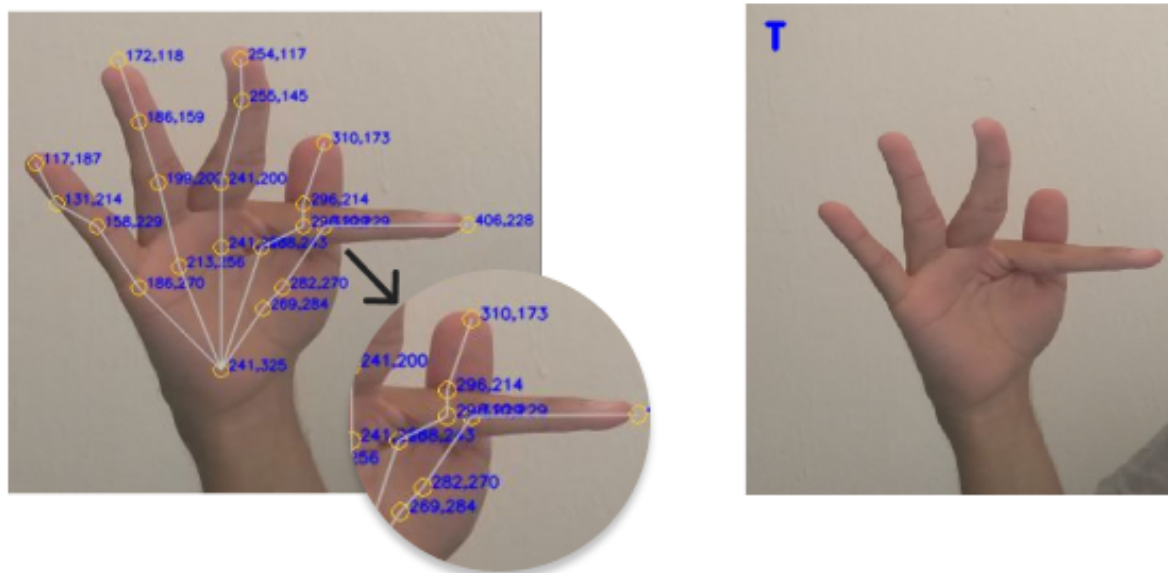
Figura 29 – Representação da (a) mão em configuração de F e (b) mão em configuração de T



FONTE: Elaborado pela autora.

A Figura 29 demonstra os sinais das letras F e T evidenciando a diferença do posicionamento do polegar na execução dos sinais.

Figura 30 – Mão em configuração de F com ligação incorreta de pontos nos dedos polegar e indicador a esquerda com círculo expondo a região de mapeamento incorreto. À direita resultado obtido.

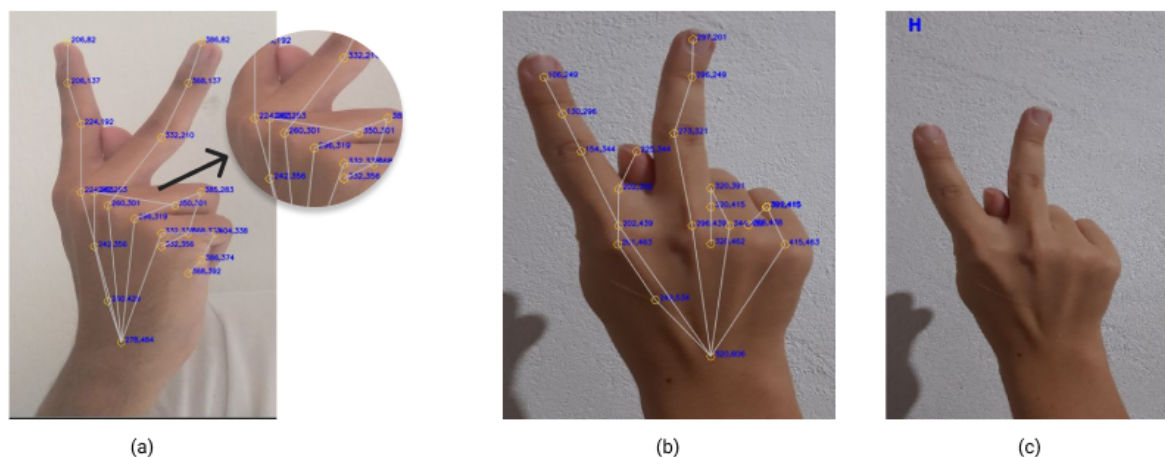


FONTE: Elaborado pela autora.

A Figura 30 mostra o sinal de F sendo executado, porém o esqueleto formado com a ligação das coordenadas do dedo polegar e indicador não foi mapeado corretamente. O algoritmo entende, por exemplo, que as coordenadas (406, 228) representam a localização do ponto 4 que é da extremidade do dedo polegar, porém o dedo que possui o ponto com essas coordenadas na Figura 29 é o indicador. Dessa forma, a característica resultante da análise do algoritmo difere da que já está definida na lista de letras do alfabeto, por isso o algoritmo não relaciona as características atribuídas a uma letra falhando na identificação do sinal. Nesse caso, a identificação obteve o resultado da tradução como letra T onde deveria ser letra F.

Em relação a letra H cuja acurácia foi de 80%, uma de suas amostras não foi traduzida devido a presença de uma característica que difere daquelas pré-definidas para o sinal da letra no algoritmo. A Figura 31 mostra o sinal com os pontos-chave mapeados onde o modelo apresentou dificuldade para localizar os pontos do polegar, que estão aproximados no círculo em destaque. Nesse caso, o algoritmo considerou como característica que o polegar não estava próximo ao dedo indicador, pois a característica retornada foi de polegar dobrado afastado do indicador.

Figura 31 – Mapeamento dos pontos do dedo polegar na amostra não identificada do sinal H.



FONTE: Elaborado pela autora.

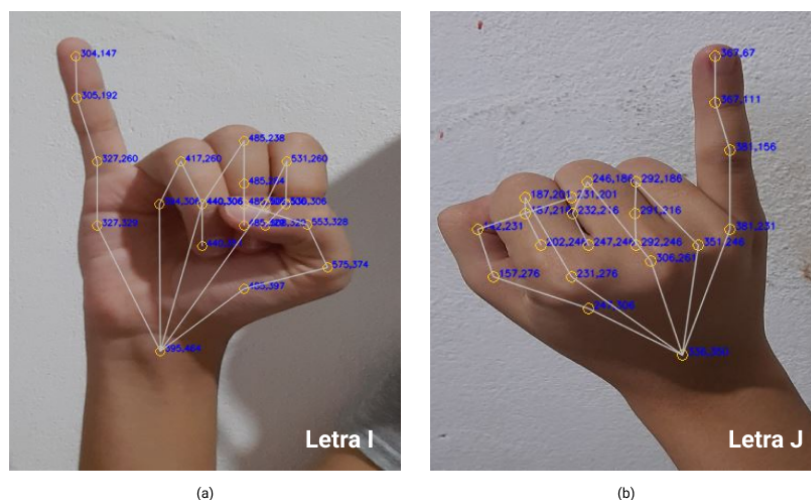
A Figura 31 apresenta em (a) o sinal da letra H com mapeamento incorreto dos pontos do polegar destacados no círculo lateral à primeira imagem, (b) apresenta mapeamento das articulações da mão em configuração de H e em (c) o resultado da análise algorítmica com resultado correto.

As características são resultantes da análise feita pelo algoritmo sobre a imagem com o cálculo da distância entre pontos e altura, além da posição da mão em relação ao pulso para serem comparadas. Na Figura 31 especificamente, a dificuldade recaiu sobre o dedo polegar.

Em relação ao sinal da letra J cuja tradução atingiu 80% de acurácia, destaca-se a igualdade com a configuração de mão do sinal da letra I, conforme demonstrado pela Figura 32, com o acréscimo do movimento circular voltando-se a palma da mão para a pessoa que o executa e seu lado contrário para a segunda pessoa da conversa. Nesse contexto, a Figura 33 mostra o resultado do algoritmo que identificou uma das amostras do sinal da letra J como sendo a letra I.

Para diferenciar uma letra da outra, optou-se por utilizar a característica do polegar onde, para a letra I, a característica seria esticado horizontal e para J seria esticado vertical, entretanto, o fato da posição da mão com o sinal de J estar ao contrário da posição da mão com o sinal de I aumenta a dificuldade de fotografar a mão de forma que o dedo polegar possa ser mapeado corretamente.

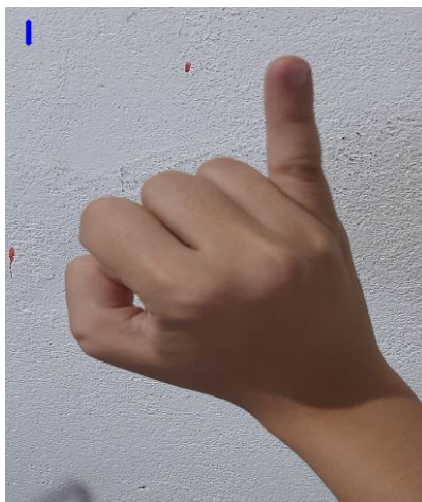
Figura 32 – Representação do (a) esqueleto do sinal da letra I e (b) esqueleto do sinal da letra J.



FONTE: Elaborado pela autora.

A Figura 32 mostra as configurações de mão das letras I e J com pontos mapeados e esqueleto formado destacando a sua igualdade. A diferença que pode-se notar na execução dos dois sinais é o fato da palma da mão estar virada ou não para quem os executa.

Figura 33 – Resultado da amostra do sinal da letra J como letra I.

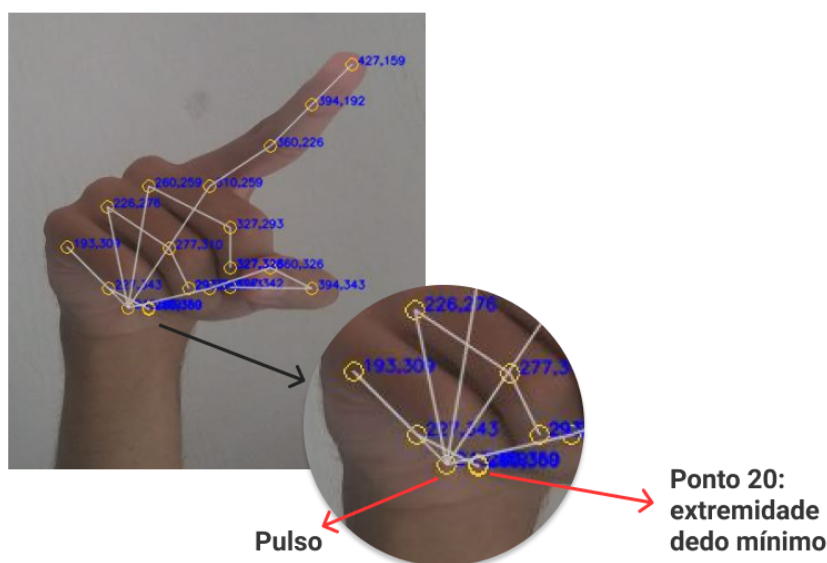


FONTE: Elaborado pela autora.

A Figura 33 demonstra a amostra cuja tradução obteve como resultado a letra I onde deveria ser letra J.

Para o sinal da letra Z, cuja acurácia foi de 80%, a amostra exposta na Figura 34 não obteve resultado, ou seja, as características obtidas a partir da análise do algoritmo não equivaliam aos conjuntos de características pré-definidos para as letras do alfabeto no algoritmo, portanto não foi possível fazer a relação para se chegar a um resultado.

Figura 34 – Amostra não identificada com mão em configuração de Z com detalhes para o ponto do pulso e ponto 20 de extremidade do dedo mínimo



que o sinal de Z possa ser traduzido pela plataforma.

Já o sinal da letra X, cuja acurácia foi de 40%, possui a configuração de mão com o dedo indicador dobrado e movimento variado dependendo de quem o executa. Para aplicar os testes da plataforma a esse sinal, definiu-se uma posição para a mão onde o indicador ficaria para cima. As amostras do sinal de X que não obtiveram o resultado correspondente a essa letra foram identificadas como letra A, conforme apresentado na Figura 36, onde a única característica que difere uma letra da outra é: se o dedo indicador está próximo ao dedo médio, então o sinal corresponde a letra A, se estiver afastado do mesmo, então o sinal corresponde a letra X, de acordo com as características pré-definidas demonstradas na Figura 37.

Figura 36 – Amostras da letra X com tradução inadequada



FONTE: Elaborado pela autora.

A Figura 36 apresenta as amostras com configuração de mão do sinal da letra X, cujo diferencial com a letra A, dada como resultado, é a proximidade do dedo indicador ao dedo médio uma vez que o dedo indicador também encontra-se dobrado na configuração do sinal da letra A. Para definir essa característica, o algoritmo subtrai o valor do eixo Y da coordenada que representa a ponta do dedo médio com o valor do eixo Y da coordenada que representa a ponta do dedo indicador da forma como é apresentado na Figura 38. Como parâmetro, definiu-se o valor 120 empiricamente para, caso o resultado da subtração seja maior que esse valor, o dedo indicador estará afastado, caso contrário, ele estará próximo ao dedo médio.

Figura 37 – Características dos sinais das letras A e X pré-definidas no algoritmo.

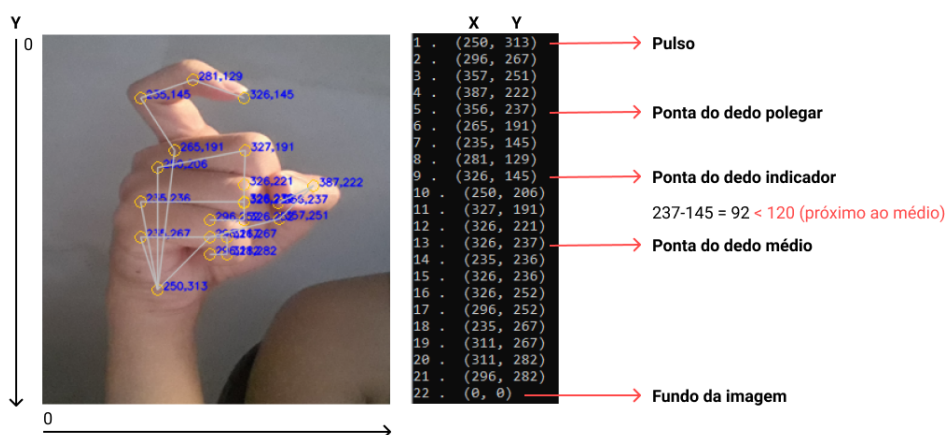
```
A = ['polegar esticado vertical: afastado do indicador',
      'indicador dobrado: proximo ao medio',
      'medio dobrado: proximo ao anelar',
      'anelar dobrado: proximo ao minimo',
      'minimo dobrado: proximo ao anelar',
      'acima']

X = ['polegar esticado vertical: afastado do indicador',
      'indicador dobrado: afastado do medio',
      'medio dobrado: proximo ao anelar',
      'anelar dobrado: proximo ao minimo',
      'minimo dobrado: proximo ao anelar',
      'acima']
```

FONTE: Elaborado pela autora.

A Figura 37 mostra as características pré-definidas para os sinais das letras A e X do algoritmo, omitindo as demais, que são usadas para que o mesmo relacione o resultado obtido em sua análise a uma letra do alfabeto já caracterizada chegando a tradução do sinal da amostra.

Figura 38 – Representação de cálculo com coordenadas do eixo Y da ponta do dedo médio e indicador.



FONTE: Elaborado pela autora.

A Figura 38 exemplifica, usando uma das amostras de X, como é realizado o cálculo citado anteriormente. Ela traz todos os valores das coordenadas dos pontos-chave mapeados destacando o pulso, ponta do dedo polegar, ponta do dedo indicador, ponta do dedo médio e fundo da imagem, e então mostra a subtração do valor de Y da coordenada da ponta do dedo médio (237) e da ponta do dedo indicador (145) com resultado menor que 120, portanto a característica retornada é "próximo ao médio" e corresponde a A segundo Figura 37.

5.3 COMPARATIVO COM OS TRABALHOS RELACIONADOS

Diante das propostas apresentadas no capítulo dos trabalhos relacionados, evidencia-se que há vários meios para desenvolver aplicações que possuem a capacidade de identificar os sinais pertencentes a LIBRAS em português, seja classificando letras ou números ou quaisquer sinais estáticos ou dinâmicos.

Nesse contexto, a Articuli traz uma abordagem que difere daquelas apresentadas pelos trabalhos utilizando a estimativa da pose humana aplicada a mão. A estimativa para a pose é feita a partir do mapeamento das articulações desse membro para obter sua localização através das coordenadas X e Y na fotografia.

Para testar a metodologia, um conjunto com 130 amostras com tamanho e ângulos variados foi montado com a participação voluntária de pessoas que não praticavam ou não tinham conhecimento em LIBRAS, com exceção de 3 pessoas que já estudaram ou possuíam alguma vivência familiar. Portanto, alguns sinais possuíam formas consideradas inexatas e mesmo assim o algoritmo conseguiu chegar a um resultado, ao contrário do que foi exposto no trabalho de

(RIBEIRO; RODRIGUES, 2017) onde o resultado é proporcional ao nível de conhecimento da pessoa que testou.

Em relação ao uso de dispositivos para captura das fotografias, a ArticuLi utiliza apenas o *smartphone* do usuário sem a necessidade de gasto para adesão de qualquer aparelho que não faça parte da vida cotidiana dos brasileiros, ao contrário, por exemplo, do LMC utilizado no estudo de (LEAL, 2018) que é de alto custo de aquisição e incomum para a maioria.

A metodologia apresentada pela ArticuLi realiza a tradução dos sinais estáticos e dinâmicos do alfabeto da LIBRAS abrangendo as letras H, K e P, cuja identificação da configuração da mão não obteve êxito com uso do LMC no trabalho de (LEAL, 2018) e as vogais identificadas no estudo de (CAVALCANTE et al., 2019), mas não trabalha com identificação de números em LIBRAS como feito por (SANTOS et al., 2020). Além disso, diferente do estudo de (MACHADO et al., 2018), a metodologia utiliza imagens ao invés de vídeos visando evitar tanto esforço repetitivo no tratamento de frames quanto computacional para o análise do vídeo.

Visto isso, a Figura 39 ilustra uma tabela que elenca as características dos trabalhos relacionados detalhando os dispositivos de captura adotados, classificadores utilizados, tipo de *dataset* e quantidade de amostras, usadas em cada trabalho.

Figura 39 – Representação comparativa entre os trabalhos relacionados e a plataforma ArticuLi detalhando dispositivos de captura, classificadores, tipo de *dataset* e quantidade de amostras

Título	Dispositivo de Captura	Classificador	Dataset	Quantidade de amostras
Tradução de Libras por Imagens	Webcam	Cascata de classificadores (<i>Machine Learning</i>)	Próprio	260 imagens
SingApp: Um Modelo de Identificação de Língua de Sinais através de Captura de Movimento em Tempo Real	Leap Motion Controller	Rede Neural Perceptron Multicamadas	Próprio	342.000 – configuração de mão e 1.800 - movimento
Classificação Automática de Sinais Visuais da Língua Brasileira de Sinais Representados por Caracterização Espaço-Temporal	Câmeras (especificações não informadas)	Rede Neural Profunda	Próprio	3.528 vídeos
Sistema Embarcado Linux para Reconhecimento de Vogais em Libras utilizando Aprendizagem de Máquina	Câmera Pi	Single Shot Multibox Detector (SSD)	Próprio	240 imagens
Um Módulo de Software para o Reconhecimento de Sinais Estáticos da Língua Brasileira de Sinais	Câmera RGB Comum	SpeededUP Robust Features (SURF)	Próprio	2.000 imagens
ArticuLi – Proposta de Tradução de Datilologia com Base em Visão Computacional para Formação de Palavras	Câmera de Smartphone	Rede Neural Convolucional	Próprio	130 imagens

FONTE: Elaborado pela autora.

Entre os trabalhos apresentados na Figura 38, destaca-se o estudo de (MACHADO et al., 2018) por utilizar o LMC, de alto custo de aquisição, mas possuir quantidade considerável de 342.000 amostras de configurações de mão e 1.800 de movimento. Nota-se também que todos utilizaram redes neurais como classificadores e bancos de amostras próprios devido a dificuldade de encontrar uma base de imagem específica da LIBRAS.

Já a Figura 40 ilustra uma tabela que põe em evidência os diferenciais da metodologia proposta por este trabalho e os demais estudos.

Figura 40 – Comparativo de características da metodologia proposta e trabalhos relacionados para evidenciar diferenciais

Título	Sinais Estáticos	Sinais Dinâmicos	Formação de Palavra	Interface	Responsividade
Tradução de Libras por Imagens	Sim	Sim	Não	Não	Não
SingApp: Um Modelo de Identificação de Língua de Sinais através de Captura de Movimento em Tempo Real	Sim	Sim	Não	Sim	Não
Classificação Automática de Sinais Visuais da Língua Brasileira de Sinais Representados por Caracterização Espaço-Temporal	Sim	Sim	Não	Não	Não
Sistema Embarcado Linux para Reconhecimento de Vogais em Libras utilizando Aprendizagem de Máquina	Sim	Não	Não	Não	Não
Um Módulo de Software para o Reconhecimento de Sinais Estáticos da Língua Brasileira de Sinais	Sim	Não	Não	Não	Não
ArticuLi – Proposta de Tradução de Datilologia com Base em Visão Computacional para Formação de Palavras	Sim	Sim	Sim	Sim	Sim

FONTE: Elaborado pela autora.

A tabela ilustrada na Figura 40 elenca a identificação de sinais estáticos do alfabeto, também realizado pelos outros estudos, sinais dinâmicos, onde os estudos de (CAVALCANTE et al., 2019) e (SANTOS et al., 2020) não se aplicam, a formação de palavra, a qual apenas a ArticuLi realiza entre os trabalhos, a presença de interface para possibilitar a interação do usuário final sendo adotada apenas no trabalho de (LEAL, 2018), mas para desktop diferente da ArticuLi, que utiliza interface *web* e com qualidade responsiva visando atender as necessidades de adequação para dispositivos com diferentes tamanhos de telas.

6 CONCLUSÃO

A comunicação é um dos fatores que proporciona inclusão em sociedade e interação entre públicos diversos, por isso tem impacto direto no cotidiano dos deficientes auditivos que, muitas vezes, não conseguem interagir com ouvintes que desconhecem sua língua e vice-versa.

Diante disso, este trabalho apresentou uma plataforma *web* intitulada Articuli cujo objetivo é atenuar ruídos sofridos durante a interação entre pessoas com deficiência auditiva e ouvintes. Ela visa estimular a interação entre esses dois grupos atuando como meio facilitador, além de proporcionar a clareza de informações de forma que as pessoas envolvidas na conversa se entendam. Além disso, a transmissão da palavra traduzida em áudio viabiliza o repasse rápido da mensagem que se quer comunicar.

Com base nesses objetivos, o foco da plataforma é realizar a tradução de sinais que pertencem ao alfabeto manual da LIBRAS em sequência para retornar a palavra relacionada em português. A plataforma localiza os pontos-chaves que representam as articulações da mão nas imagens de entrada através de um modelo pré-treinado com base no *dataset* da COCO e realiza análise algorítmica com uso da biblioteca OpenCV.

Os testes utilizando a metodologia proposta foram realizados com imagens fotografadas por 10 voluntários usando seus *smartphones* como dispositivo de captura. Com isso, obteve-se um conjunto balanceado com 5 amostras para cada sinal totalizando 130 imagens. O conjunto está armazenado no github da autora e pode ser acessado através do endereço eletrônico <https://github.com/andressamesquita/ImagensLibrasParaTeste>.

Como resultados, 21 sinais alcançaram resultados corretos em 100% das amostras, são eles: A, B, C, D, E, G, I, K, L, M, N, O, P, Q, R, S, T, U, V, W e Y. Já as letras F, H, J e Z chegaram a 80% e X a 40% de acurácia. Com isso, das 130 imagens analisadas pela ferramenta, 123 foram traduzidas com sucesso chegando a 94,6% de acurácia pela plataforma de modo geral, tornando-a apta a formar palavras atendendo a seus objetivos.

6.1 TRABALHOS FUTUROS

Diante das características presentes nas fotografias capturadas pelos voluntários que podem interferir na análise realizada pelo algoritmo, sugere-se como trabalho futuro implementar o processamento de imagem para tratar as sombras que se apresentam próximas aos dedos das mãos na execução dos sinais.

Já em relação aos testes, a montagem do conjunto de fotografias foi feita a partir da colaboração de voluntários, assim sugere-se como trabalho futuro testar a metodologia proposta com uma base de imagens maior de sinais do alfabeto da LIBRAS.

REFERÊNCIAS

- ADORF, J. Web speech api. *KTH Royal Institute of Technology*, 2013. Citado na página 26.
- BARBOSA, H. J. C. et al. Perfil clínico epidemiológico de pacientes com perda auditiva. *Journal of Health & Biological Sciences*, v. 6, n. 4, p. 424–430, 2018. Citado na página 17.
- BENASSI, C. A.; PADILHA, S. de J. Fonologia da libras os parâmetros e a relação pares mínimos na libras. *Revista Diálogos*, v. 3, n. 2, p. 94–106, 2016. Citado na página 18.
- BICAS, R. d. S.; GUIJO, L. M.; DELGADO-PINHEIRO, E. M. C. Habilidades auditivas e de comunicação oral de crianças e adolescentes deficientes auditivos e o processo de reabilitação fonoaudiológica. *Revista CEFAC*, SciELO Brasil, v. 19, n. 4, p. 465–474, 2017. Citado na página 17.
- BONATTO, A. N. Desenvolvimento de protótipo para reconhecimento de gestos de uma das mãos. 2018. Citado 2 vezes nas páginas 18 e 19.
- BRASIL. *LEI Nº 10.436, DE 24 DE ABRIL DE 2002*. 2002. <http://www.planalto.gov.br/ccivil_03/LEIS/2002/L10436.htm>. Online; accessed 29 november 2018. Citado 2 vezes nas páginas 14 e 18.
- CARDOSO, N. S.; BISPO, T. M. d. S. et al. Um estudo comparativo entre os principais frameworks de desenvolvimento web em linguagem python. UFRA/Campus Belém (PA), 2019. Citado na página 26.
- CAVALCANTE, G. S. et al. Sistema embarcado linux para reconhecimento de vogais em libras utilizando aprendizagem de máquina. In: SBC. *Anais da VII Escola Regional de Computação Aplicada à Saúde*. [S.l.], 2019. p. 235–240. Citado 3 vezes nas páginas 31, 53 e 54.
- COCOLO, C. *Reconhecimento de movimentos de cães utilizando um acelerômetro e redes neurais artificiais*. Tese (Doutorado) — Universidade de São Paulo, 2020. Citado 2 vezes nas páginas 22 e 23.
- CORRÊA, Y. et al. Produção científica brasileira sobre tradução automática português brasileiro-libras: uma revisão sistemática de literatura. *Revista virtual de estudos da linguagem-ReVEL*. Novo Hamburgo, RS. Vol. 16, nesp 15 (nov. 2018), p. 179-203, 2018. Citado na página 14.
- DAI, H. et al. Joint coco and mapillary workshop at iccv 2019 keypoint detection challenge track technical report: Distribution-aware coordinate representation for human pose estimation. *arXiv preprint arXiv:2003.07232*, 2020. Citado na página 20.
- DIAS, T. S. et al. *Luva instrumentada para reconhecimento de padrões de gestos em Libras*. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2020. Citado na página 28.
- FARIA, E. L. d. Redes neurais convolucionais e máquinas de aprendizado extremo aplicadas ao mercado financeiro brasileiro. Universidade Federal do Rio de Janeiro, 2018. Citado na página 24.

- FERNANDES, C. C.; ROMEIRO, C. de A. A contribuição da datilologia como estratégia metodológica no processo de alfabetização. *Revista diálogos*, v. 4, n. 1, p. 170–185, 2017. Citado na página 18.
- FERREIRA, W. F. d. Um sistema para organização de acervos baseado em visão computacional. Universidade Federal do Pampa, 2020. Citado na página 20.
- FONSÊCA, R. O. d.; DUTRA, M. R. P.; FERREIRA, M. Â. F. Satisfação de usuários com aparelhos de amplificação sonora individual concedidos pelo sistema único de saúde: revisão integrativa. *Audiology-Communication Research*, SciELO Brasil, v. 25, 2020. Citado na página 18.
- IBGE. *Censo Demográfico 2010*. 2010. Online; accessed 27 september 2020. Citado na página 14.
- IQBAL, U. et al. Hand pose estimation via latent 2.5 d heatmap regression. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. p. 118–134. Citado 2 vezes nas páginas 21 e 22.
- JARDIM, D. S.; MACIEL, F. J.; LEMOS, S. M. A. Perda auditiva incapacitante: análise de fatores associados. *Audiology-Communication Research*, SciELO Brasil, v. 22, 2017. Citado na página 17.
- LEAL, M. M. Singapp: um modelo de identificação de língua de sinais através de captura de movimento em tempo real. Universidade do Vale do Rio dos Sinos, 2018. Citado 4 vezes nas páginas 14, 29, 53 e 54.
- LIN, T.-Y. et al. *Microsoft COCO: Common objects in context*. 2015. Arxiv. Disponível em: <<https://arxiv.org/abs/1405.0312>>. Citado na página 25.
- MACHADO, M. C. et al. Classificação automática de sinais visuais da língua brasileira de sinais representados por caracterização espaço-temporal. Universidade Federal do Amazonas, 2018. Citado 4 vezes nas páginas 19, 30, 53 e 54.
- MALHEIROS, I. S. e M. VisualizaÇÃo em tempo real da estrutura esquelética de uma mão humana usando realidade aumentada. *Revista Destaques Acadêmicos*, v. 11, n. 4, 2019. ISSN 2176-3070. Disponível em: <<http://univates.br/revistas/index.php/destaques/article/view/2409>>. Citado na página 39.
- NEWELL, A.; YANG, K.; DENG, J. Stacked hourglass networks for human pose estimation. In: SPRINGER. *European conference on computer vision*. [S.l.], 2016. p. 483–499. Citado 2 vezes nas páginas 7 e 21.
- OLIVEIRA, B. R. d. et al. Modelagem e desenvolvimento de dispositivo para reabilitação de dedos da mão. Universidade Federal de Goiás, 2016. Citado na página 39.
- PINHEIRO, A.; SILVA, T. L. da. A importância da visão computacional e suas aplicações na mobilidade urbana. In: *XII Mostra de Iniciação Científica e Extensão Comunitária e XI Mostra de Pesquisa de Pós-Graduação IMED 2018*. [S.l.: s.n.], 2018. Citado na página 19.
- RIBEIRO; RODRIGUES, D. R. S. M. T. A. N. Tradução de libras por imagem. Universidade de Itauna, 2017. Citado 2 vezes nas páginas 28 e 53.

- RODRIGUES, A.; ALMEIDA-SILVA, A. Reflexões sociolinguísticas sobre a libras (língua brasileira de sinais). *Estudos Linguísticos (São Paulo)*, v. 46, n. 2, p. 686–698, 2017. Citado na página 18.
- RODRIGUES, N. R. P. et al. *Deteção e caracterização de pose humana*. Dissertação (Mestrado), 2018. Citado na página 14.
- ROSA, I. d. M. F. et al. Mapeamento estruturado da libras para utilização em sistemas de comunicação. *Sistema Maxwell / Laboratório de Automação de Museus, Bibliotecas Digitais e Arquivos*, n. 48, 2016. Citado na página 18.
- ROSA, R. d. P. et al. Método de classificação de pragas por meio de rede neural convolucional profunda. Universidade Estadual de Ponta Grossa, 2018. Citado na página 23.
- ROSTIANINGSIH, S.; SETIAWAN, A.; HALIM, C. I. Coco (creating common object in context) dataset for chemistry apparatus. *Procedia Computer Science*, Elsevier, v. 171, p. 2445–2452, 2020. Citado na página 14.
- SANTOS, A. V. et al. Rede neural artificial convolucional aplicada ao reconhecimento de configuração de mão nos símbolos de 0 a 9 da língua brasileira de sinais (libras). In: *Anais Estendidos do XV Simpósio Brasileiro de Sistemas de Informação*. Porto Alegre, RS, Brasil: SBC, 2019. p. 21–24. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbsi_estendido/article/view/7432>. Citado na página 22.
- SANTOS, A. V. et al. Um módulo de software para o reconhecimento de sinais estáticos da língua brasileira de sinais. *Revista Acta Kariri-Pesquisa e Desenvolvimento*, v. 2, n. 1, 2020. Citado 4 vezes nas páginas 14, 31, 53 e 54.
- SILVA, L. M. G. Detecção de colônias de bactérias utilizando visão computacional. Pontifícia Universidade Católica de Goiás, 2020. Citado na página 20.
- SILVA, R. E. V. d. Um estudo comparativo entre redes neurais convolucionais para a classificação de imagens. 2018. Citado na página 24.
- SILVA, R. O. da; SILVA, I. R. S. Linguagem de programação python. *TECNOLOGIAS EM PROJEÇÃO*, v. 10, n. 1, p. 55–71, 2019. Citado na página 26.
- SILVA, S. d. M.; FARINELLI, F.; ALMEIDA, M. B. Um roteiro para modelagem conceitual de sistemas de informação baseada em princípios ontológicos. 2017. Citado na página 33.
- SOLIQUE, I. de S.; SASS, G. G.; LIMA, M. A. C. Estudo comparativo das tecnologias de desenvolvimento de sistemas web: Polymer, django e meteor. UEMS, 2018. Citado na página 26.
- SOUZA, L. P.; SANTO, F. do E. Comparativo entre frameworks de css bootstrap e bulma para desenvolvimento de projetos web. *Revista Interface Tecnológica*, v. 17, n. 1, p. 140–152, 2020. Citado na página 26.
- SUN, K. et al. Deep high-resolution representation learning for human pose estimation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2019. p. 5693–5703. Citado na página 21.

WETTERICH, C. B.; BARROSO, H. C. S. M.; FREITAS, D. A. A comunicação entre surdos e profissionais da saúde: Uma revisão bibliográfica. *Educação Profissional e Tecnológica em Revista*, v. 4, n. 1, p. 130–152, 2020. Citado na página 15.

WHO. *Sordera y pérdida de la audición*. 2019. Online; accessed 26 september 2020. Citado 2 vezes nas páginas 14 e 15.

ZHENG, C. et al. *Deep Learning-Based Human Pose Estimation: A Survey*. 2021. Citado 3 vezes nas páginas 20, 22 e 25.