MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Especialização em Ciência de Dados

Aplicação de técnicas de reconhecimento de imagens na classificação de sinais em LIBRAS (Linguagem brasileira de sinais) para tradução em texto

Balbino Soares Ferreira e Victor Silva Grossi

João Monlevade, MG
2024

VICTOR SILVA GROSSI
BALBINO SOARES FERREIRA

APLICAÇÃO DE TÉCNICAS DE RECONHECIMENTO DE IMAGENS NA CLASSIFICAÇÃO DE SINAIS EM LIBRAS (LINGUAGEM BRASILEIRA DE SINAIS) PARA TRADUÇÃO EM TEXTO

Trabalho de Conclusão de curso apresentado como requisito parcial para obtenção do título de Especialização em Ciências de dados na Universidade Federal de Ouro Preto.

Orientador(a): Prof. Dr. Matheus Nohra Haddad

JOÃO MONLEVADE

2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

G878a Grossi, Victor Silva.
Aplicação de técnicas de reconhecimento de imagens na classificação de sinais em LIBRAS (Linguagem brasileira de sinais) para tradução em texto. [manuscrito] / Victor Silva Grossi. Balbino Soares Ferreira. - 2024.
45 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Matheus Haddad.
Produção Científica (Especialização). Universidade Federal de Ouro Preto. Departamento de Engenharia de Produção.

1. Classificação - Sinais e símbolos. 2. Estatística matemática. 3. Língua brasileira de sinais. 4. Língua de sinais. 5. Redes neurais (Computação). 6. Sistemas de reconhecimento de padrões - Imagens digitais. 7. Tradução e interpretação. I. Ferreira, Balbino Soares. II. Haddad, Matheus. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 519.237.8:004.8

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Balbino Soares Ferreira Filho

Victor Silva Grossi

Aplicação de técnicas de reconhecimento de imagens na classificação de sinais em LIBRAS (Linguagem brasileira de sinais) para tradução em texto

Trabalho de conclusão de curso apresentado ao curso de Especialização em Ciência de Dados da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Especialista em Ciência de Dados

Aprovada em 04 de julho de 2024

Membros da banca

Prof. Dr. Matheus Nohra Haddad - Orientador (Universidade Federal de Ouro Preto)
Prof. Dr. Pablo Luiz Araujo Munhoz (Universidade Federal de Ouro Preto)
Prof. Dr. Harlei Miguel de Arruda Leite (Instituto Tecnológico de Aeronáutica)

Matheus Nohra Haddad, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 14/08/2024

Documento assinado eletronicamente por **Matheus Nohra Haddad**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 14/08/2024, às 17:57, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0759626** e o código CRC **87DA1608**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0819 - www.ufop.br

Resumo

O reconhecimento de sinais em Libras é um processo que apresenta grandes desafios, não somente com relação ao problema da classificação de imagens e vídeos, mas como uma lingua de natureza visual-gestual, apresentando os mesmos dificuldades de problemas envolvendo linguagem natural . Neste contexto, temos ainda agregado o fato que a lingua brasileira de sinais possuir poucos estudos de pesquisa. Este trabalho possui como proposta avaliar modelos de classificação de vídeos que possam ser aplicadas no processo de classificação de sinais em libras e determinar o modelo que apresente melhor performance. Foi utilizada uma base de dados com vinte sinais de libras, gravada por 12 sinalizadores 5 vezes cada para o processo de treinamento. Por fim avaliamos os principais modelos de extração de características utilizando modelos pré-treinados e uma técnica de *Deep Learning* para classificar os sinais baseado nos acertos de predição realizada por uma rede Neural Convulacional. Foi avaliada os modelos CNN+RNN e *Transformers*, onde foi possível comparar os resultados alçados entre eles, de forma a validar o modelo com melhor resultados na base treinada.

Palavras-chaves: *Deep Learning*. Redes Neurais Convulacionais. Libras. Reconhecimento automático de sinais.

Abstract

Recognizing signs in Libras is a process that presents significant challenges, not only concerning the problem of image and video classification but also because it is a visually-gestural language, presenting the same difficulties as problems involving natural language. In this context, we also face the fact that there are few research studies on Brazilian Sign Language. This work aims to evaluate video classification models that can be applied to the process of classifying Libras signs and determine the model that shows the best performance. A dataset with twenty Libras signs, recorded by 12 signers, each five times, was used for the training process. Finally, we evaluated the main feature extraction models using pre-trained models and a Deep Learning technique to classify the signs based on the prediction accuracy achieved by a Convolutional Neural Network. The CNN+RNN and Transformers models were evaluated, allowing us to compare the results obtained between them, in order to validate the model with the best results on the trained dataset.

Keywords: Deep Learning. Convulational Neural Network. Libras. Automatic Signs Recognition.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CNN Redes Neurais convulacionais

CNN-3D Redes Neurais Convolucionais 3D

GAN *Generative Adversarial Networks*

GPU Unidade de Processamento Gráfico

GRU *Gated Recurrent Unit*

LFS Lingua Francesa de Sinais

LIBRAS Lingua Brasileira de Sinais

LSTM *Long Short Term Memory*

ONU Organização das Nações Unidas

PLN Processamento de Linguagem Natural

RGB *Red, Green, Blue*

RGB-D *Red, Green, Blue - Depth*

RNN Redes Neurais Recorrentes

ST-GCN Redes Neurais Temporais e Espaciais

Sumário

1 Introdução

Embora a busca por uma sociedade inclusiva e integrada não seja algo novo, somente nas últimas três décadas vem ganhando papel de destaque em políticas governamentais tanto nacionais quanto no âmbito global. O conceito de sociedade inclusiva aparece pela primeira vez na resolução geral número 45/91 da Organização das Nações Unidas (ONU). Esta resolução apresentou o termo sociedade inclusiva, alterando a abordagem do programa das nações unidas para deficientes de conscientização para ações mais concretas. (SILVA *et al.*, 2015)

As primeiras políticas voltadas para pessoas com deficiência, tinham como objetivo apenas abrigar e alimentar, mantendo estas pessoas em instituições, onde eram segregadas da sociedade. Posteriormente avançamos para uma fase de integração destas pessoas com a sociedade. E mais recentemente temos as políticas de inclusão. Enquanto na integração apenas a pessoa com deficiência se adapta para integrar a sociedade, a inclusão visa qualificar a sociedade para que possa incluir a todos. Na sociedade inclusiva, todos se adaptam e se transformam para que as necessidades e diferenças de cada um sejam respeitadas e consideradas, permitindo as mesmas oportunidades a todos. Além disso não somente promover esta inclusão, devendo evitar a exclusão. (HAZARD; FILHO; REZENDE, 2007).

No Brasil, o tema vem gradativamente sendo incorporado na legislação. Já na constituição de 1967, com a regra genérica de igualdade, bem como em posterior emenda constitucional que tratava garantir aos deficientes a igualdade, a acessibilidade, além da integração social. O tema foi mantido no texto constitucional de 1988, que trata a questão da deficiência centralizando o tema no princípio da igualdade. Além da matriz da igualdade formal, ou seja, todos são iguais perante a lei, encontraremos a regra da igualdade material, ou seja, o suporte dado pelo Estado, reconhecendo situação de vulnerabilidade de determinados grupos (ARAUJO, 2008).

No tema relativo à inclusão de pessoas com deficiência auditiva e de fala, temos a Lei nº 10.436 de 2002 que é a lei reconhecendo a LIBRAS como meio legal de comunicação e expressão no país. Recentemente temos a lei brasileira de inclusão (Lei nº 13.146/2015), na qual as pessoas surdas são contempladas em vários aspectos, entre eles: direito a educação bilíngue, ensino de LIBRAS, formação de tradutores e intérpretes de LIBRAS, tradução de editais em LIBRAS, sendo alguns direitos adquiridos e outros reafirmados por meio de tal legislação (ROCHA; PASIAN, 2023).

Embora a língua brasileira de sinais seja um mecanismo que possibilite a inclusão de pessoas com deficiências auditivas e de fala na sociedade, ainda são encontradas barreiras principalmente pela dificuldade de comunicação entre as pessoas surdas que tem domínio de LIBRAS e as pessoas que não têm conhecimento nesta linguagem. Outro ponto que chama a atenção é que muitas pessoas com deficiência geralmente fazem uso da leitura orofacial e estabelecem a comunicação via língua portuguesa mesmo que de forma precária devido à dificuldade em se comunicar (ROCHA; PASIAN, 2023).

A troca de informações é fundamental para a comunicação humana, utilizando os sentidos sensoriais para expressar ideias e estabelecer diálogos por meio das estruturas linguísticas (SMITH, 2010). Enquanto a comunicação verbal é comum, a comunicação por sinais se destaca, especialmente nas línguas de sinais, que utilizam movimentos manuais, corporais e expressões faciais para se comunicar com pessoas surdas (JONES; PATEL, 2015). No Brasil, LIBRAS é reconhecida oficialmente desde 2002 pela Lei nº 10.436 (**??**) e está em constante evolução, especialmente em seu vocabulário (SILVA, 2018). A regulamentação da LIBRAS em 2005 pelo Decreto nº 5.626 (BRASIL, 2005) abriu caminho para debates sobre inclusão e acessibilidade para surdos.

Para facilitar a comunicação, a utilização de tecnologias emergentes, particularmente a inteligência artificial e a visão computacional, tem mostrado grande potencial. A inteligência artificial, por meio de técnicas de aprendizado de máquina, permite o desenvolvimento de sistemas que podem interpretar e gerar sinais em língua de sinais com maior precisão. De acordo com (LI; WU; ZHANG, 2020), algoritmos de inteligência artificial podem aprender e generalizar padrões de sinais a partir de grandes volumes de dados, melhorando a eficácia da comunicação. Além disso, a visão computacional é fundamental para a análise de sinais em vídeos e imagens. Técnicas de processamento de imagem e redes neurais convolucionais são utilizadas para reconhecer e classificar sinais visuais (SIMONYAN; ZISSERMAN, 2014a), possibilitando a interpretação automática e a integração desses sinais em interfaces interativas.

Assim como a inteligência artificial tem avançado na tradução entre diferentes idiomas, facilitando a comunicação global entre pessoas de diversas culturas e nacionalidades, a visão computacional pode desempenhar um papel crucial na melhoria da comunicação para indivíduos com deficiência auditiva e de fala. Utilizando essa tecnologia, é possível desenvolver sistemas que ajudem na tradução e interpretação da LIBRAS, permitindo que pessoas que não dominam essa língua possam interagir de maneira mais eficiente com aqueles que a utilizam.

Apesar das leis que respaldam a LIBRAS, a comunidade surda enfrenta desafios na comunicação devido à falta de preparo e conhecimento dos ouvintes em geral (RODRIGUES, 2008). Muitos não conhecem a língua ou acreditam em uma língua de sinais universal, o que dificulta a disseminação da língua na sociedade.

De maneira geral, os sinais em LIBRAS são representados por uma combinação de gestos manuais, movimentos corporais e expressões faciais, formando uma linguagem rica e multifacetada. Isso introduz uma complexidade adicional ao processo de classificação, pois é necessário analisar não apenas um sinal de forma isolada, mas o contexto completo analisando as expressões, os sinais e a sequência de movimento corporal, para assim realizar uma classificação precisa e eficaz. A tarefa de identificar e interpretar sinais dentro de sequências de vídeo exige que o sistema compreenda a interação entre diferentes elementos visuais e temporais.

O treinamento de modelos de aprendizado de máquina é um processo crucial para a classificação automática de sinais em LIBRAS. Durante o treinamento, um modelo computacional é exposto a um conjunto de dados rotulado, no qual cada exemplo de entrada é associado a uma saída conhecida. O objetivo é ajustar os parâmetros do modelo para minimizar a diferença entre as previsões feitas pelo modelo e as saídas reais. Esse ajuste é realizado através de algoritmos de otimização, que iterativamente ajustam os parâmetros com base no erro calculado entre a previsão e a verdade conhecida.

No contexto da visão computacional e reconhecimento de sinais, o treinamento envolve a exposição do modelo a uma grande quantidade de dados de vídeo, onde cada vídeo contém sinais que precisam ser identificados e classificados. O modelo aprende a identificar padrões e características distintivas dos sinais ao longo de várias iterações. O treinamento é essencial para que o modelo adquira a capacidade de generalizar para novos dados, ou seja, ser capaz de reconhecer sinais que não foram explicitamente vistos durante o treinamento.

A ausência de uma base de dados abrangente e bem estruturada para LIBRAS é um desafio significativo para o desenvolvimento de sistemas de reconhecimento. Embora existam diversos esforços na construção de tais bases, elas ainda são pequenas em comparação com a estimativa de cerca de 14 mil sinais conhecidos. Esta limitação de dados impede a criação de modelos robustos e generalizáveis.

Além disso, o treinamento de modelos para a classificação de sinais em LIBRAS requer uma capacidade computacional substancial. O treinamento de modelos de aprendizado profundo, frequentemente utilizados para essas tarefas, exigem recursos computacionais avançados, como unidades de processamento gráfico e grandes volumes de dados para ajustar e validar os parâmetros dos modelos. O processamento de vídeos, que envolve a análise de sequências extensas de imagens, amplifica ainda mais a demanda por poder computacional. Portanto, superar esses desafios é crucial para o desenvolvimento eficaz de sistemas automáticos de reconhecimento de sinais em LIBRAS.

Primeiramente, os modelos de aprendizado profundo, que são frequentemente utilizados para tarefas de reconhecimento de sinais, demandam considerável poder computacional. Esses modelos, como redes neurais convolucionais e redes neurais recorrentes, são projetados para identificar e interpretar padrões complexos em dados visuais e temporais. O treinamento desses modelos exige unidades de processamento gráfico avançadas, que são capazes de realizar cálculos intensivos de maneira eficiente e em paralelo. As GPUs são essenciais para acelerar o processo de treinamento, permitindo que o modelo ajuste seus parâmetros com base em grandes quantidades de dados.

O processamento de vídeos adiciona uma camada adicional de complexidade ao cenário. Cada vídeo é composto por uma sequência de imagens, e a análise dessas sequências envolve a identificação e interpretação de padrões temporais além dos padrões espaciais. Esse aspecto temporal amplifica a demanda computacional, uma vez que o modelo precisa considerar a dinâmica dos sinais ao longo do tempo, e não apenas em uma única imagem estática.

Diante desses desafios, é fundamental desenvolver estratégias eficazes para otimizar o treinamento dos modelos. Isso pode incluir técnicas como a redução de dimensionalidade dos dados, uso de pré-processamento eficiente e implementação de métodos de treinamento distribuído para maximizar o uso dos recursos computacionais disponíveis. Superar essas dificuldades é essencial para criar sistemas automáticos de reconhecimento de sinais em LIBRAS que sejam precisos, robustos e capazes de operar em cenários reais e diversos.

1.1 Objetivo geral

O objetivo geral deste trabalho é contribuir com estudos na área de inteligência artificial e visão computacional de forma a auxiliar na inclusão de pessoas com deficiência auditiva e de fala, através do estudo de modelos de classificação de vídeos, identificar uma metodologia e base de dados que possa ser utilizada no treinamento dos modelos, utilizar e comparar modelos de extração de características através dos modelos pré-treinados Inception V3, DenseNet121 e VGG16 e aplicar e comparar os modelos de classificação CNN+RNN e Inception V3.

1.1.1 Objetivos específicos

Para atingir o objetivo geral é necessário atender aos seguintes objetivos específicos:

- Identificar os principais modelos utilizados na classificação de vídeo
- Identificar uma base de dados com sinais em LIBRAS
- Validar uma metodologia de treinamento e classificação.
- Determinar o modelo com melhor performance;

1.2 Contribuições

O trabalho oferece contribuições significativas no campo do reconhecimento de sinais em LIBRAS destacando o potencial das atividades computacionais nesse contexto, através de técnicas de aprendizado de máquina e visão computacional. Além disso, fornece uma visão abrangente sobre a utilidade prática da metodologia aplicada, indicando seu potencial para melhorar a inclusão social e a comunicação para pessoas com deficiência auditiva e de fala.

1.3 Organização do Trabalho

O restante deste trabalho é organizado como segue: o Capítulo 2 apresenta os detalhes dos modelos e o referencial teórico. O capítulo 3 apresenta o formato que foi definido o problema. O capítulo 4 descreve a metodologia aplicada na execução do trabalho. O Capítulo 5 apresenta e discute os resultados. As considerações finais estão no Capítulo 6.

2 Referencial Teórico

A integração de tecnologias e metodologias para o processamento da LIBRAS requer uma abordagem multidisciplinar que envolve diversos conceitos avançados. A LIBRAS, enquanto uma forma de comunicação visual-espacial, apresenta características linguísticas e estruturais próprias que demandam técnicas especializadas para sua análise e interpretação.

O reconhecimento de sinais é um processo que envolve a conversão de informações visuais em dados que possam ser compreendidos e manipulados por sistemas computacionais. Esse processo é realizado por meio da aplicação de algoritmos e métodos que permitem a identificação e a interpretação dos sinais de forma precisa e eficiente.

As CNN têm se mostrado particularmente eficazes para o reconhecimento de padrões visuais, sendo uma ferramenta fundamental na análise de imagens e vídeos. Estas redes são projetadas para lidar com a complexidade dos dados visuais e são essenciais para a construção de sistemas que possam realizar a tarefa de reconhecimento de sinais com alta precisão.

A classificação de vídeos é uma etapa importante no processamento de sinais, envolvendo técnicas de aprendizado de máquina para categorizar e interpretar sequências visuais. Essa etapa é crucial para a correta identificação e interpretação dos sinais em vídeos, considerando padrões temporais e espaciais.

A utilização de bases de dados adequadas é indispensável para o desenvolvimento e a validação de modelos de reconhecimento. Essas bases fornecem os dados necessários para o treinamento dos modelos, possibilitando o aprimoramento e a adaptação dos sistemas para lidar com uma variedade de sinais e contextos.

Este referencial teórico oferece uma visão abrangente sobre a LIBRAS e as tecnologias associadas ao seu processamento, destacando a importância da combinação de técnicas avançadas e dados relevantes para a criação de sistemas de reconhecimento de sinais eficientes.

2.1 Língua Brasileira de Sinais

No Brasil, o surgimento da língua de sinais remonta ao Segundo Império, quando o educador francês Hernest Hurt chegou ao país a pedido de Dom Pedro II, que tinha um neto surdo (DUARTE, 2013). Com ele, veio o alfabeto manual francês e a Lingua Francesa de Sinais (LFS), dando origem à LIBRAS, influenciada principalmente pela LFS.

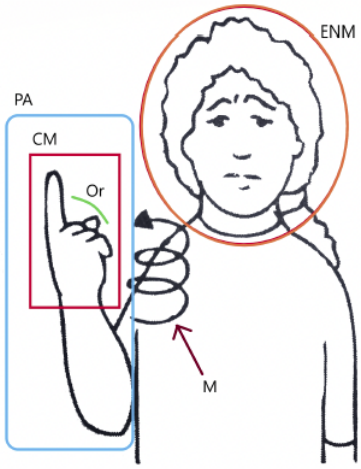

Figura 1 – Parâmetros fonológicos de LIBRAS. Fonte:(REZENDE; SILVA; GUIMARãES, 2021)

O reconhecimento oficial da LIBRAS veio com a Lei nº 10.436/2002 e sua regulamentação pelo Decreto nº 5.626/2005 (BRASIL, 2005), representando um avanço significativo para a comunidade surda, garantindo-lhes direitos e apoio para sua integração na sociedade. A lei define a LIBRAS como um sistema linguístico completo, com gramática própria, essencial para a comunicação e expressão das comunidades surdas do Brasil.

A compreensão da LIBRAS como uma linguagem natural, com estrutura linguística independente da língua oral, é fundamental para o processo de reconhecimento e tradução dos sinais. Ferreira-Brito (1993) destaca a distinção entre línguas de sinais e mímica, enfatizando a complexidade linguística das primeiras. GESSER, Audrei. (2009) complementa, ressaltando que, embora as línguas de sinais possam usar representações icônicas, há sinais arbitrários, convencionados pela comunidade surda. A menor unidade significativa das línguas de sinais é o sinal, caracterizado por parâmetros como ponto de articulação (PA), orientação das palmas das mãos (Or), configuração das mãos (CM), movimento (M) e expressões não-manuais (STOKOE, 1960). Conforme pode ser visualizado na Figura 1.

O aprendizado da LIBRAS segue os princípios de aprendizado de línguas, variando conforme a idade, motivação e contexto do aluno. A imersão na cultura surda é essencial para fluência, embora não haja um número fixo de sinais para alcançá-la (CAPOVILLA; MARTINS; OLIVEIRA, 2017). Cada língua de sinais é única, com suas próprias características culturais, apesar das influências mútuas entre elas (Simons, G. and Fennig, C., 2018).

Desta forma a Língua Brasileira de Sinais, conforme definido pela Lei de LIBRAS de 2002, é um sistema de comunicação de natureza visual-motora, com estrutura gramatical própria, usado na transmissão de ideias e fatos através de gestos, expressões faciais e corporais (STELLE; STRIEICHEN, 2013).

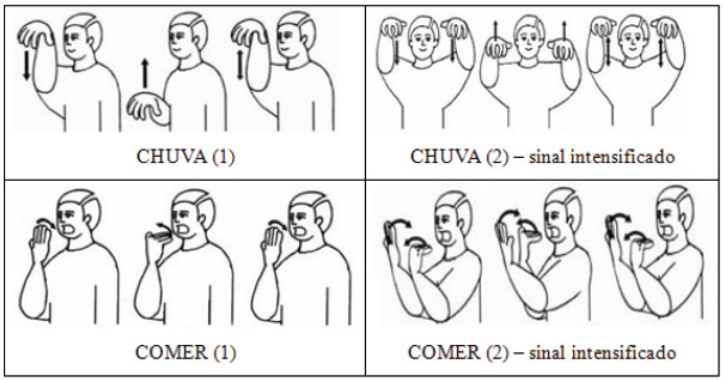

Figura 2 – Sinais intensificados em libras: Fonte (FERREIRA *et al.*, 2018)

Como nas linguas escritas, a LIBRAS também possui um alfabeto manual, que representa letras e números. Porém a comunicação em línguas de sinais ocorre principalmente através de sinais específicos e não apenas pela tradução das letras para o alfabeto manual. A literatura demonstra que a comunicação em LIBRAS utiliza sinais específicos para a maioria das palavras O alfabeto manual somente é utilizado em casos em que ainda não foi convencionado um sinal para representar aquela palavra pela comunidade. (STELLE; STRIEICHEN, 2013).

Enquanto na comunicação em línguas orais o foco geralmente está nas palavras e sons, na língua de sinais outros articuladores além das mãos são utilizados, como outras partes do corpo, face e tronco. Onde as palavras são representadas por sinais e os sons por movimento. Os cinco parâmetros formativos usados na linguagem são (SOUZA TANYA A. FELIPE DE; MONTEIRO, 2007):

1. A configuração da mão;
2. Ponto ou local de articulação;
3. O movimento;
4. Orientação/direcionalidade;
5. Expressão facial e/ou corporal.

Segundo (FERREIRA *et al.*, 2018) as expressões não manuais, que podem variar desde movimentos do corpo a expressões faciais desempenham um importante papel para expressar intensidade, ênfase, concordância e interrogativa. Conforme demonstrado na Figura 2.

2.2 Reconhecimento de Sinais em LIBRAS

O reconhecimento de sinais é uma área de pesquisa em visão computacional e aprendizado de máquina que visa identificar e interpretar gestos humanos a partir de dados visuais. Essa disciplina tem despertado um interesse significativo, como evidenciado por uma série de estudos realizados ao longo das décadas. Desde o trabalho pioneiro de Stokoe (1960), que estabeleceu os parâmetros básicos das línguas de sinais, até os avanços recentes, como o desenvolvimento de sistemas de reconhecimento de gestos baseados em inteligência artificial.

O reconhecimento de sinais em LIBRAS emprega uma variedade de técnicas de processamento de sinais que permitem transformar gestos e movimentos em dados compreensíveis para sistemas computacionais. Entre as abordagens mais comuns estão a visão computacional, redes neurais convolucionais e sensores avançados.

A visão computacional é uma das técnicas predominantes, utilizando imagens ou vídeos dos sinais para análise. Em estudos como o de (SANTOS; COSTA; SILVA, 2018), intitulado "Reconhecimento de LIBRAS usando dados de profundidade e cor", a combinação de dados de profundidade e cor é usada para aprimorar a precisão do reconhecimento dos sinais. O processamento de sinais nesse contexto inclui a extração de características das imagens, como os contornos e movimentos das mãos, seguido pela aplicação de algoritmos de aprendizado de máquina para classificar os sinais. Técnicas como extração de características, segmentação de imagem e análise de movimento são essenciais para identificar e interpretar os sinais de LIBRAS.

As CNNs também desempenham um papel importante no reconhecimento de se sinais em LIBRAS. No trabalho de (NGUYEN; NGUYEN; PHAM, 2020), "Reconhecimento de Língua de Sinais Usando Técnicas de Aprendizado Profundo", as CNNs são empregadas para reconhecer sinais em vídeos. Essas redes neurais são projetadas para processar dados com uma estrutura de grade, como imagens, e são utilizadas para extrair características hierárquicas dos sinais. O processamento de sinais aqui envolve a transformação de dados visuais em representações que as CNNs podem analisar, utilizando técnicas como *pooling* e normalização para melhorar a precisão do reconhecimento.

Outra abordagem significativa é o uso de sensores no corpo, que capturam os movimentos das mãos e braços. O estudo de (CHEN; YANG; HSU, 2017), "Reconhecimento de Língua de Sinais Baseado em Sensores *Wearables* Usando Modelo de Markov Oculto", explora a aplicação de sensores vestíveis para captar os sinais. O processamento de sinais nesta abordagem envolve a interpretação dos dados dos sensores por meio de Modelos Ocultos de Markov, que são eficazes para reconhecer padrões temporais e sequenciais. Técnicas de fusão de dados e análise temporal são utilizadas para interpretar os sinais baseando-se nas variações e dinâmicas dos dados capturados.

Além disso, o rastreamento de movimento tridimensional é utilizado em tecnologias como o *Kinect.* O estudo de (KIM; CHOI; LEE, 2019), "Reconhecimento em Tempo Real da Língua de Sinais Usando Kinect", demonstra como dados tridimensionais dos movimentos das mãos e braços são processados para reconhecer sinais em tempo real. Técnicas como mapeamento espacial, rastreamento de articulações e análise de postura são aplicadas para converter os dados tridimensionais em informações úteis para o reconhecimento de sinais.

Essas técnicas, quando combinadas com algoritmos avançados de aprendizado de máquina e *hardware* especializado, formam a base dos sistemas modernos de reconhecimento de LIBRAS. Elas permitem uma interpretação mais precisa dos sinais, contribuindo para uma comunicação mais eficiente e inclusiva para a comunidade surda. A evolução contínua nessas áreas é fundamental para aprimorar a precisão e a acessibilidade desses sistemas.

Um marco importante na evolução do reconhecimento de sinais foi a introdução de técnicas de pré-processamento de imagens, como destacado por HASAN, M. and MISRA, M. (2012). Essas técnicas visam remover ruídos e outros elementos irrelevantes das imagens, preparando-as para a extração de características. A extração de características, como mencionado por Escalera, Sergio e Vassilis Athitsos e Guyon (2016), é uma etapa crucial que converte a informação visual em uma forma numérica compreensível pelo computador, permitindo análises mais sofisticadas.

No campo da aquisição de dados, houve uma diversificação significativa de abordagens ao longo do tempo. Autores como Starner, Weaver e Pentland (1997) e Rao e Kishore (2017) exploraram diferentes métodos de captura de sinais, desde sistemas acoplados à cabeça do sinalizador até o uso de câmeras de smartphones. Essas técnicas variadas refletem a necessidade de adaptar a aquisição de dados às especificidades de cada projeto de pesquisa.

No entanto, um desafio persistente enfrentado pelos pesquisadores é a falta de bases de dados padronizadas e abrangentes para treinar e testar os sistemas de reconhecimento. Muitos estudos ainda criam suas próprias bases de dados, o que leva a conjuntos de dados fragmentados e limitados em escala. A superação desse obstáculo exige um esforço colaborativo da comunidade científica para desenvolver conjuntos de dados representativos e escalonáveis (KHAN; AWAIS; QURESHI, 2022).

Além disso, a evolução das ferramentas de Visão Computacional, como o OpenCV (OPENCV, 2024), tem impulsionado o desenvolvimento de sistemas de reconhecimento de sinais. Essas ferramentas permitem o processamento eficiente de imagens, incluindo a remoção de ruídos, a detecção de regiões de interesse e a extração de características relevantes, como destacado por VIEVILLE e CRAHAY (2004) e ASSALEH e AL-ROUSAN (2005).

Por fim, é importante mencionar a crescente integração de informações faciais nos sistemas de reconhecimento de sinais. Autores como INFANTINO, RIZZO e GAGLIO (2007) e Rao e Kishore (2017) exploraram o uso de expressões faciais como parte dos parâmetros fonológicos das línguas de sinais, ampliando as possibilidades de identificação e interpretação dos gestos.

Apesar do interesse crescente na comunidade de aprendizado de máquina e visão computacional no reconhecimento de gestos *(Gesture Recognition)* (ESCALERA; SERGIO; VASSILIS ATHITSOS E GUYON, 2016), a classificação de sinais em LIBRAS ainda é pouco explorada em comparação com outras línguas de sinais. A falta de bases de dados é um fator significativo que limita o avanço nesta área, resultando na criação de bases de dados próprias por muitos trabalhos (REZENDE, 2021).

2.3 Redes Neurais Artificiais

Redes neurais artificiais têm se tornado um campo de pesquisa fundamental em inteligência artificial, com aplicações que vão desde a visão computacional até o processamento de linguagem natural. As CNN têm se destacado especialmente em tarefas relacionadas ao reconhecimento de imagens e vídeos devido à sua capacidade de extrair características hierárquicas e complexas dos dados (LECUN; BENGIO; HINTON, 2015). Essas redes são compostas por camadas convolucionais que aplicam filtros para capturar padrões locais, camadas de pooling que reduzem a dimensionalidade e camadas totalmente conectadas que realizam a classificação (GOODFELLOW; BENGIO; COURVILLE, 2016).

O desenvolvimento das CNNs foi um marco importante na evolução das redes neurais, com o surgimento de arquiteturas inovadoras como a *AlexNet*, que demonstrou a eficácia das CNNs em competições de reconhecimento de imagem (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). A capacidade destas redes de aprender representações detalhadas de dados visuais tem levado a avanços significativos em diversas áreas, incluindo a detecção de objetos e a segmentação semântica.

Além disso, redes neurais profundas têm sido amplamente utilizadas para abordar problemas complexos em diferentes domínios. Por exemplo, (MNIH *et al.*, 2015) demonstraram como redes neurais profundas podem ser aplicadas ao aprendizado por reforço, alcançando desempenho semelhante ao de humanos em jogos de eletrônicos mais antigos. Outro aspecto relevante é o estudo das leis de escalabilidade das redes neurais profundas, que revelam como o desempenho das redes pode ser aprimorado com o aumento da quantidade de dados e da profundidade da rede (HESTNESS *et al.*, 2017).

Além das aplicações diretas, as redes neurais têm contribuído para o desenvolvimento de métodos explicáveis em inteligência artificial, permitindo que modelos complexos sejam mais interpretáveis e compreensíveis para os usuários (CARUANA *et al.*, 2015; RIBEIRO; SINGH; GUESTRIN, 2016). Essas abordagens são cruciais para a adoção prática e segura de redes neurais em setores sensíveis, como saúde e finanças.

Em resumo, o campo das redes neurais continua a evoluir rapidamente, com novas técnicas e descobertas ampliando constantemente suas aplicações e potencial. As pesquisas atuais se concentram em melhorar a eficiência e a aplicabilidade dos modelos, bem como em explorar novas áreas de aplicação e desafios (BENGIO, 2019; VOULODIMOS *et al.*, 2018).

2.4 Rede Neural Convolucional (CNN)

O uso de redes neurais convolucionais (CNN) para o reconhecimento da língua brasileira de sinais (LIBRAS) é uma abordagem popular e eficaz devido à capacidade das CNNs de extrair características espaciais hierárquicas de imagens e vídeos. Essa técnica se baseia na aplicação de camadas convolucionais para identificar padrões e características em dados visuais, o que é particularmente útil para interpretar sinais de LIBRAS capturados em imagens ou vídeos.

As CNNs são projetadas para processar dados com uma estrutura de grade, como imagens, e são capazes de aprender representações hierárquicas dos sinais. Por exemplo, o estudo de (NGUYEN; NGUYEN; PHAM, 2020), explora a utilização de CNNs para o reconhecimento de sinais em vídeos. Neste trabalho, as CNNs são empregadas para analisar frames de vídeos de sinais, extraindo características como formas e movimentos das mãos. A arquitetura desta rede permite que o sistema identifique e classifique sinais de LIBRAS com alta precisão, utilizando camadas convolucionais para capturar padrões complexos e camadas de pooling para reduzir a dimensionalidade e melhorar a eficiência computacional.

Além das CNNs, outras arquiteturas de redes neurais têm sido aplicadas ao reconhecimento de LIBRAS, cada uma trazendo vantagens específicas para diferentes aspectos do problema. Entre essas arquiteturas, destacam-se:

- RNN: são especialmente úteis para tarefas que envolvem dados sequenciais, como vídeos de sinais de LIBRAS. As RNNs podem modelar a dinâmica temporal dos sinais, capturando a sequência de movimentos ao longo do tempo. Um estudo relevante é o de (SARMA; ZHANG; LIU, 2020), que investiga a combinação de RNNs com CNNs para melhorar o reconhecimento de sinais de LIBRAS. Neste trabalho, as CNNs são usadas para extrair características espaciais das imagens, enquanto as RNNs modelam as relações temporais entre os sinais em sequência, resultando em um sistema de reconhecimento mais robusto e preciso.

- CNN-3D: expandem a ideia das CNNs para trabalhar com vídeos em vez de imagens estáticas, capturando informações espaciais e temporais simultaneamente. O estudo de (KWON; KIM; CHO, 2019), demonstra como estas redes podem ser aplicadas para reconhecer sinais de LIBRAS em vídeos. Elas aplicam convoluções ao longo do tempo e espaço, permitindo que o modelo aprenda tanto as características visuais quanto as dinâmicas temporais dos sinais, o que é crucial para uma interpretação precisa.

- Redes Neurais de Atenção (*Attention Networks*): incluindo a arquitetura *Transformer*, são capazes de focar em partes relevantes dos dados e ignorar informações irrelevantes. O trabalho de (XU; ZHANG; CHEN, 2021), explora como a arquitetura *Transformer* pode ser aplicada ao reconhecimento de LIBRAS. O *Transformer* usa mecanismos de atenção para destacar partes importantes dos sinais e melhorar a precisão do reconhecimento ao lidar com informações contextuais complexas.

Essas abordagens mostram como diferentes arquiteturas de redes neurais podem ser utilizadas para abordar o problema do reconhecimento de LIBRAS, cada uma oferecendo características únicas que contribuem para a eficácia dos sistemas de tradução de sinais. A escolha da arquitetura depende do tipo de dados, da complexidade dos sinais e das necessidades específicas do sistema de reconhecimento.

O emprego das CNN tem sido objeto de inúmeras pesquisas e estudos, destacando-se como uma das ferramentas mais poderosas em visão computacional e aprendizado profundo. Autores como LeCun *et al.* (1998) foram pioneiros ao introduzir a arquitetura *LeNet*, que revolucionou o reconhecimento de padrões em imagens digitais. Posteriormente, Krizhevsky, Sutskever e Hinton (2012) apresentaram a *AlexNet*, uma CNN profunda que conquistou o primeiro lugar no desafio *ImageNet*, impulsionando a popularização e o desenvolvimento contínuo dessa tecnologia.

Avançando para o cenário atual, pesquisas como as de He *et al.* (2016) sobre a *ResNet* e SZEGEDY *et al.* (2015) sobre a *InceptionNet* trouxeram contribuições significativas para o campo, abordando desafios como a dissipação do gradiente em redes profundas e a eficiência computacional. Essas arquiteturas estabeleceram novos padrões de desempenho em tarefas de classificação de imagens, além de influenciar o design de outras redes neurais convolucionais.

Além disso, a aplicação das CNN se estende além do domínio do reconhecimento de imagens. Estudos como o de Long, Shelhamer e Darrell (2015) exploraram as CNN em tarefas de segmentação semântica, enquanto o trabalho de Szegedy, Ioffe e Vanhoucke (2016) introduziu a ideia de redes neurais residuais para melhorar a aprendizagem em profundidade. Essas abordagens demonstraram sucesso em diversas áreas, incluindo medicina, análise de vídeo, reconhecimento de voz e muito mais.

No entanto, é fundamental reconhecer que o avanço contínuo das redes neurais convolucionais CNNs não é impulsionado exclusivamente por inovações arquitetônicas, mas também por melhorias significativas na coleta e disponibilidade de conjuntos de dados rotulados. O impacto desses conjuntos de dados é profundo, pois eles fornecem a base essencial para o treinamento e a avaliação eficaz dos modelos de CNN.

Pesquisadores desempenharam um papel crucial no desenvolvimento e disponibilização de conjuntos de dados massivos e diversificados, que têm sido fundamentais para a evolução das CNNs. Um exemplo notável é o conjunto de dados *ImageNet*, criado por (DENG *et al.*, 2009). A *ImageNet* é um banco de dados abrangente que contém milhões de imagens rotuladas, abrangendo milhares de categorias. Este conjunto de dados não apenas facilitou o treinamento de redes neurais em larga escala, mas também estabeleceu *benchmarks* importantes para a avaliação de desempenho dos modelos.

Outro exemplo relevante é o conjunto de dados *Microsoft COCO,* desenvolvido por (LIN *et al.*, 2014). O COCO inclui imagens ricas em conteúdo e anotações detalhadas que abrangem várias tarefas de visão computacional, como detecção de objetos e segmentação semântica. A diversidade e a complexidade dos dados presentes no COCO têm sido essenciais para o treinamento de modelos de CNN e para a condução de avanços em tarefas de visão computacional.

A disponibilidade de conjuntos de dados de grande escala e de alta qualidade é, portanto, um fator determinante no progresso das CNNs. Eles possibilitam que as redes neurais sejam treinadas com um número vasto de exemplos, melhorando sua capacidade de generalização e sua resistência ao overfitting. Essa base de dados robusta permite também a comparação e a validação de diferentes abordagens e técnicas, promovendo um avanço mais rápido e fundamentado no campo das redes neurais convolucionais.

As CNN representam uma ferramenta essencial no arsenal de técnicas de aprendizado profundo, com uma base sólida de pesquisas que abrangem várias décadas e uma ampla gama de aplicações. O contínuo avanço nesse campo promete trazer ainda mais inovações e impactos significativos em nossa sociedade digitalizada.

A arquitetura das CNNs é projetada de maneira a imitar a organização do córtex visual humano, onde neurônios individuais respondem a estímulos em campos receptivos locais, permitindo uma representação eficiente e hierárquica de características visuais (LECUN *et al.*, 1998). Essa inspiração biológica proporciona uma base sólida para a capacidade das CNNs de aprender representações discriminativas a partir de dados brutos, sem a necessidade de extração manual de características.

Uma das contribuições mais significativas dos modelos de CNN é sua capacidade de generalização. Por meio do treinamento em grandes conjuntos de dados rotulados, esses modelos podem extrapolar padrões aprendidos para dados não vistos, permitindo uma ampla gama de aplicações no mundo real (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).A capacidade de generalização das redes neurais convolucionais (CNNs) é particularmente importante em cenários onde a disponibilidade de dados é limitada, como em tarefas médicas de diagnóstico por imagem. A habilidade das CNNs de aprender características gerais a partir de um conjunto restrito de dados é essencial para a criação de modelos eficazes que podem ser aplicados a novos dados e contextos. A relevância dessa capacidade é bem documentada em estudos recentes sobre a aplicação de deep learning em análise de imagens médicas (LITJENS *et al.*, 2017).

Os avanços na arquitetura das CNN levaram ao desenvolvimento de modelos cada vez mais complexos e eficientes. Inovações como redes residuais (*ResNets*), que introduzem conexões de atalho para facilitar o treinamento de redes profundas, e redes neurais densamente conectadas *DenseNets*, que promovem um fluxo de informação mais eficiente entre as camadas, têm impulsionado os limites do desempenho em tarefas desafiadoras, como classificação de imagens de alta resolução e segmentação semântica (HE *et al.*, 2016; HUANG *et al.*, 2017).

Além disso, a transferência de aprendizado *(transfer learning)* emergiu como uma estratégia poderosa para aproveitar o conhecimento adquirido por modelos pré-treinados em conjuntos de dados massivos, como o *ImageNet*, e adaptá-los para tarefas específicas com conjuntos de dados menores (YOSINSKI *et al.*, 2014). essa abordagem reduz a necessidade de grandes volumes de dados de treinamento e pode acelerar significativamente o desenvolvimento de soluções em diversos domínios. Técnicas como o aprendizado por transferência *(transfer learning)* têm demonstrado como é possível utilizar modelos pré-treinados para melhorar o desempenho em tarefas específicas com conjuntos de dados limitados, facilitando o desenvolvimento de soluções em áreas diversas, como reconhecimento de imagem e processamento de linguagem natural (PAN; YANG, 2010).

No entanto, mesmo com todos os avanços, os modelos de CNN ainda enfrentam desafios significativos. A interpretabilidade dos modelos continua sendo uma área ativa de pesquisa, com esforços para entender como as decisões são tomadas e identificar possíveis vieses ou falhas nos algoritmos (SAMEK *et al.*, 2017). Questões éticas relacionadas à privacidade e ao uso responsável da tecnologia também estão ganhando destaque, à medida que os modelos de CNN são cada vez mais integrados em sistemas autônomos e de tomada de decisão (HOLSTEIN; VAUGHAN; III, 2019).

Os modelos de classificadores em (CNN) representam uma conquista notável na interseção entre inteligência artificial e visão computacional. Sua capacidade de aprender representações complexas de dados visuais impulsionou avanços significativos em uma variedade de aplicações, desde diagnósticos médicos precisos até veículos autônomos seguros. À medida que continuamos a desbravar novos horizontes na pesquisa em (CNN), é essencial considerar não apenas o potencial desses modelos, mas também as questões éticas e sociais que surgem com seu uso disseminado. A busca por sistemas de inteligência artificial justos, transparentes e responsáveis é fundamental para garantir que os benefícios dessas tecnologias sejam equitativamente distribuídos e sustentáveis a longo prazo (ESTEVA *et al.*, 2017; BOJARSKI *et al.*, 2016).

Nos últimos anos, os modelos de redes neurais convolucionais CNNs têm desempenhado um papel crucial no avanço da visão computacional e do processamento de imagens. Esses modelos se destacam por sua capacidade de aprender representações hierárquicas de dados visuais através de camadas de convolução e *pooling*. As camadas de convolução são responsáveis por extrair características locais das imagens, aplicando filtros convolucionais que capturam padrões e texturas. Já as camadas de *pooling* reduzem a dimensionalidade dos dados processados, agregando informações e mantendo as características mais relevantes enquanto diminuem a complexidade computacional (GOODFELLOW; BENGIO; COURVILLE, 2016).

Entre esses modelos, o Inception V3 e o VGG16 destacam-se como marcos importantes na história da pesquisa em redes neurais convolucionais. O Inception V3 foi introduzido pelo Google Research como parte da família Inception, apresentando uma arquitetura altamente eficiente e capaz de lidar com uma ampla variedade de tarefas de visão computacional (SZEGEDY; IOFFE; VANHOUCKE, 2016). A arquitetura do Inception V3 é caracterizada por sua capacidade de processamento em várias escalas, graças aos seus módulos de convolução Inception. Esses módulos permitem a captura eficiente de características visuais em diferentes níveis de abstração, levando a uma representação mais rica e discriminativa das imagens. Essa abordagem tem sido fundamental para o sucesso do Inception V3 em várias competições de classificação de imagens e detecção de objetos (SZEGEDY; IOFFE; VANHOUCKE, 2016).

O VGG16, desenvolvido pelo Visual Geometry Group da Universidade de Oxford, é conhecido por sua simplicidade e profundidade em camadas. Com 16 camadas, incluindo convoluções e camadas totalmente conectadas, o VGG16 estabeleceu um padrão de referência para muitas tarefas de visão computacional devido à sua eficácia e facilidade de implementação (SIMONYAN; ZISSERMAN, 2014b). O VGG16 é elogiado por sua simplicidade e facilidade de entendimento, servindo como uma base sólida para a exploração e desenvolvimento de novas técnicas na área (SIMONYAN; ZISSERMAN, 2014b).

Ambos os modelos, Inception V3 e VGG16, têm sido amplamente adotados e adaptados em diversas aplicações, desde classificação de imagens até segmentação semântica e detecção de objetos. A escolha entre esses modelos geralmente depende das necessidades específicas da aplicação, dos recursos computacionais disponíveis e dos requisitos de desempenho (SZEGEDY; IOFFE; VANHOUCKE, 2016; SIMONYAN; ZISSERMAN, 2014b).

Apesar de suas diferenças arquiteturais, o Inception V3 e o VGG16 compartilham um objetivo comum: extrair representações significativas e discriminativas de dados visuais, permitindo uma ampla gama de aplicações em visão computacional e aprendizado de máquina. O contínuo desenvolvimento e aprimoramento desses modelos prometem avançar ainda mais os limites da capacidade de processamento de imagens e da inteligência artificial como um todo.

2.5 CNN+RNN

A arquitetura CNN+RNN combina convoluções para processamento espacial com camadas recorrentes para processamento temporal, utilizando uma rede CNN+ e uma rede RNN com camadas *Gated Recurrent Unit* (GRU). As camadas GRU, introduzidas por Cho *et al.* (2014), são uma variação das *Long Short Term Memory* (LSTM) e visam superar alguns problemas das RNNs.

A combinação de Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN) oferece uma integração eficiente das capacidades de extração de características espaciais e análise temporal. As CNNs são particularmente eficazes na identificação de padrões e estruturas em dados espaciais, como imagens e vídeos. Por outro lado, as RNNs são projetadas para processar e analisar sequências temporais, capturando dependências e relações de longo prazo dentro das sequências. A combinação dessas duas redes permite a extração detalhada de características espaciais e a modelagem precisa da dinâmica temporal, sendo especialmente útil em tarefas complexas como a análise de vídeos, onde é crucial entender tanto as características dos frames quanto sua evolução ao longo do tempo.

2.6 *Transformers*

Mais recentemente, em 2017, foi proposta a arquitetura *Transformer* pelos pesquisadores do Google, inicialmente para tarefas de tradução automática de texto. Essa arquitetura se mostrou eficaz em uma variedade de tarefas de Processamento de Linguagem Natural (PLN) e é adequada para classificação de vídeo, pois lida com dados sequenciais e captura dependências globais entre os frames (VASWANI *et al.*, 2017).

Além disso, a abordagem *Transformer* tem se destacado devido à sua eficiência em lidar com grandes volumes de dados e à sua capacidade de capturar relações de longo alcance entre elementos de dados por meio de mecanismos de atenção. Introduzida como uma alternativa às RNNs, a arquitetura *Transformer* é capaz de processar sequências sem depender de estruturas recorrentes, permitindo uma análise mais eficaz e escalável. Essa abordagem tem liderado o estado da arte em diversas tarefas de aprendizado de máquina, incluindo processamento de linguagem natural e visão computacional, devido à sua capacidade de generalização e ao seu desempenho superior.

2.7 Classificação de sinais de LIBRAS em vídeos

A classificação de vídeos é uma disciplina complexa que tem sido alvo de extensa pesquisa e desenvolvimento devido à sua ampla gama de aplicações em diversas áreas, como vigilância por vídeo, análise de mídia social e reconhecimento de atividades humanas (KARPATHY; FEI-FEI, 2014). Essa área se beneficiou significativamente do avanço das redes neurais convolucionais, que permitem a análise e compreensão de sequências de vídeo através da extração e interpretação de características temporais e espaciais. Ao longo dos anos, uma variedade de abordagens e técnicas foram propostas para lidar com os desafios únicos apresentados pela classificação de vídeos.

No campo da visão computacional e aprendizado profundo, as redes neurais convolucionais (CNNs) desempenharam um papel crucial no avanço da classificação de vídeos. Autores como Simonyan e Zisserman (2014b) e Carreira e Zisserman (2017) introduziram arquiteturas de redes neurais convolucionais profundas, como as CNN-3D, que são capazes de capturar informações temporais e espaciais simultaneamente, permitindo uma representação mais rica e robusta de vídeos. A utilização das CNN-3D é especialmente relevante no reconhecimento de LIBRAS, onde a capacidade de modelar a dinâmica dos sinais ao longo do tempo é crucial. O estudo de Kwon, Kim e Cho (2019) demonstra como as CNN-3D aplicam convoluções ao longo do tempo e espaço para identificar padrões complexos nos sinais, permitindo uma análise aprofundada dos movimentos das mãos e braços.

Além das CNNs, as redes neurais recorrentes (RNNs) têm sido amplamente utilizadas para modelar a estrutura temporal de vídeos. Autores como Donahue *et al.* (2016) e Venugopalan *et al.* (2015) exploraram o uso das arquiteturas LSTM e GRU para capturar padrões temporais de longo prazo em sequências de vídeo, melhorando assim a capacidade de classificação. No contexto de LIBRAS, a combinação de RNNs com CNNs tem mostrado resultados promissores. O trabalho de Sarma, Zhang e Liu (2020) combina CNNs para extração de características espaciais com RNNs para modelagem da sequência temporal dos sinais, resultando em uma abordagem robusta para o reconhecimento dos sinais de LIBRAS.

Outro avanço significativo na classificação de vídeos é o uso de Modelos de Atenção, como os *Transformers*. O estudo de Xu, Zhang e Chen (2021) explora a aplicação de *Transformers* para reconhecimento de sinais, usando mecanismos de atenção para destacar partes importantes dos vídeos e melhorar a precisão da classificação. Esses modelos ajudam a focar nas características mais relevantes dos sinais de LIBRAS e são eficazes para lidar com a complexidade e variabilidade dos dados.

Além das técnicas convencionais de aprendizado de máquina, a classificação de vídeos também se beneficia dos avanços em redes neurais temporais e espaciais, como as Redes Neurais Temporais e Espaciais (ST-GCN). O trabalho de Yan, Xiong e Lin (2018) demonstra como estas redes podem ser aplicadas para modelar a estrutura dos movimentos das mãos e braços em LIBRAS, utilizando grafos temporais para capturar a dinâmica dos sinais.

A análise de vídeos de comprimentos variados é outro aspecto importante na classificação de vídeos. Autores como Ng *et al.* (2015) e Zhu *et al.* (2016) investigaram técnicas para lidar com a variabilidade na duração dos vídeos, desenvolvendo abordagens que processam vídeos de comprimentos arbitrários de forma eficiente e escalável. Em LIBRAS, onde a duração dos sinais pode variar, essas técnicas são cruciais para garantir a precisão na interpretação dos sinais.

Além disso, técnicas como *Generative Adversarial Networks* (GAN) e *Transfer Learning* têm sido exploradas para melhorar a classificação de vídeos. Estudos como Vondrick, Pirsiavash e Torralba (2016) e Mahajan *et al.* (2018) investigaram o uso de GANs para gerar sequências realistas de *frames* de vídeo, enquanto Zhou *et al.* (2018) e Tran *et al.* (2018) exploraram o uso de *transfer learning* para adaptar modelos pré-treinados a tarefas específicas de classificação de vídeos. Essas abordagens ajudam a melhorar a robustez e a generalização dos modelos de classificação de vídeos, incluindo o reconhecimento de LIBRAS.

A classificação de vídeos é uma área de pesquisa dinâmica e em constante evolução, impulsionada por avanços em técnicas de aprendizado de máquina, disponibilidade de conjuntos de dados rotulados e poder computacional crescente. O desenvolvimento de algoritmos eficazes para classificação de vídeos tem o potencial de impactar positivamente uma variedade de aplicações práticas e continuar a moldar o futuro da inteligência artificial e da análise de mídia visual.

2.8 Base de dados de sinais em LIBRAS

As bases de dados representam o cerne da infraestrutura de informações em diversos setores, desde empresas até instituições acadêmicas. Autores como Elmasri e Navathe (2015) e Silberschatz, Korth e Sudarshan (2010) são referências incontornáveis no estudo das bases de dados, fornecendo insights essenciais sobre modelagem, projeto e implementação de sistemas de gerenciamento de banco de dados.

A evolução das tecnologias de *big data* e computação em nuvem trouxe novos desafios e oportunidades para o campo das bases de dados. Reddy, Aggarwal e Reddy (2018) e Batini *et al.* (2009) discutem a importância da qualidade dos dados, destacando a necessidade de técnicas eficazes de limpeza e validação para garantir a integridade e confiabilidade dos dados armazenados.

A segurança e privacidade dos dados emergiram como preocupações críticas na era digital. Ferraiolo, Cugini e Kuhn (2001) e Stoneburner, Goguen e Feringa (2002) abordam a importância de medidas de segurança robustas, como criptografia e controle de acesso, para proteger os dados contra ameaças cibernéticas e garantir conformidade com regulamentações de privacidade.

A análise de dados é uma aplicação fundamental das bases de dados modernas. Han, Kamber e Pei (2011) e Witten *et al.* (2016) exploram técnicas avançadas de mineração de dados e aprendizado de máquina que permitem extrair *insights* valiosos e padrões ocultos dos dados, contribuindo para previsões precisas e tomadas de decisão informadas.

Estratégias eficazes de armazenamento e indexação são essenciais para garantir a recuperação eficiente dos dados. Garcia-Molina, Ullman e Widom (2002) e Borthakur *et al.* (2007) discutem tecnologias de armazenamento de dados, como bancos de dados distribuídos e sistemas de arquivos distribuídos, que permitem escalabilidade e gerenciamento eficaz de grandes volumes de dados.

A integração de sistemas e a troca de dados entre diferentes plataformas são desafios comuns enfrentados na gestão de bases de dados. Berners-Lee, Hendler e Lassila (2001) e McIlraith, Son e Zeng (2001) investigam técnicas de integração de dados semânticos e ontologias que permitem harmonizar e unificar dados de fontes heterogêneas, facilitando a interoperabilidade e a troca de informações entre sistemas.

No contexto envolvendo o aprendizado de máquina envolvendo língua de sinais, as bases de dados específicas desempenham um papel crucial para o desenvolvimento e treinamento dos modelos. No contexto de línguas de sinais podemos destacar:

A RWTH-PHOENIX-Weather 2014T, uma base de dados de sinais da língua alemã de sinais, fornece vídeos anotados com sinais realizados por vários usuários em diferentes condições. Este recurso tem sido utilizado para treinar e avaliar modelos de reconhecimento de sinais, como demonstrado em Koller, Ney e Bowden (2016). Esta base de dados é essencial para o treinamento de Redes Neurais Convolucionais (CNNs) e Redes Neurais Recorrentes (RNNs), facilitando o reconhecimento e a interpretação de sinais ao longo do tempo.

A SLR-100K é uma base de dados de vídeos contendo sinais em *Flemish Sign Language*, ou Língua de Sinais Flamenga, que é uma língua de sinais utilizada na Bélgica, coletados em diversas condições e contextos. O estudo de Zhu, Zhang e Liu (2019) aproveita esta base de dados para investigar a eficácia de técnicas de *Transfer Learning* e redes neuronais convolucionais para melhorar a precisão do reconhecimento de sinais em vídeos. A diversidade e a quantidade de dados oferecidos por esta base são fundamentais para a construção de modelos robustos e generalizáveis.

A GS-DataSet é outra base importante que inclui sinais em *Greek Sign Language*, a linguagem de sinais utilizada na Grécia, com anotações detalhadas sobre a execução e a interpretação dos sinais. A pesquisa de Li, Liu e Xu (2018) utiliza esta base para explorar modelos de ST-GCN no reconhecimento de sinais de LIBRAS. A GS-DataSet é valiosa para a análise da estrutura e dinâmica dos sinais, permitindo o desenvolvimento de abordagens que combinam a análise espacial e temporal dos movimentos.

Além dessas, a *Sign Language Dataset for Deep Learning (SL-Deep)* oferece uma coleção diversificada de sinais em *American Sign Language*, que é a Língua de Sinais Americana, e é utilizada para avaliar técnicas avançadas como redes neurais de atenção e *Transformers*. O trabalho de Xu, Zhang e Chen (2021) demonstra como esta base é usada para melhorar a precisão dos modelos de reconhecimento por meio de mecanismos de atenção que destacam partes relevantes dos sinais.

Na literatura temos poucas bases para a língua brasileira de sinais, por este motivo a maioria dos autores optam por criar suas próprias bases de dados. Rezende (2021), propõe em seu trabalho uma base de dados e metodologia específica para Língua Brasileira de Sinais. Sendo está a base MINDS-Libras *Dataset*, composta por mais de 1100 amostras, com 20 sinais em LIBRAS gravados por doze pessoas diferentes usando sensores *Red, Green, Blue - Depth* (RGB-D) e câmeras *Red, Green, Blue* (RGB). Os sinais foram selecionados por uma profissional proficiente em LIBRAS, com base na variabilidade dos parâmetros fonológicos, como configuração da mão, ponto de articulação, movimento das mãos, orientação da palma e expressões não-manuais. Na Figura 3, podemos verificar os sinais presentes nesta base.

As bases de dados específicas para LIBRAS desempenham um papel vital na pesquisa e desenvolvimento de tecnologias para o reconhecimento automático de sinais. Elas fornecem os dados necessários para treinar e testar algoritmos de aprendizado profundo, e suas características únicas ajudam a abordar os desafios associados à interpretação dos sinais em diferentes contextos e condições.

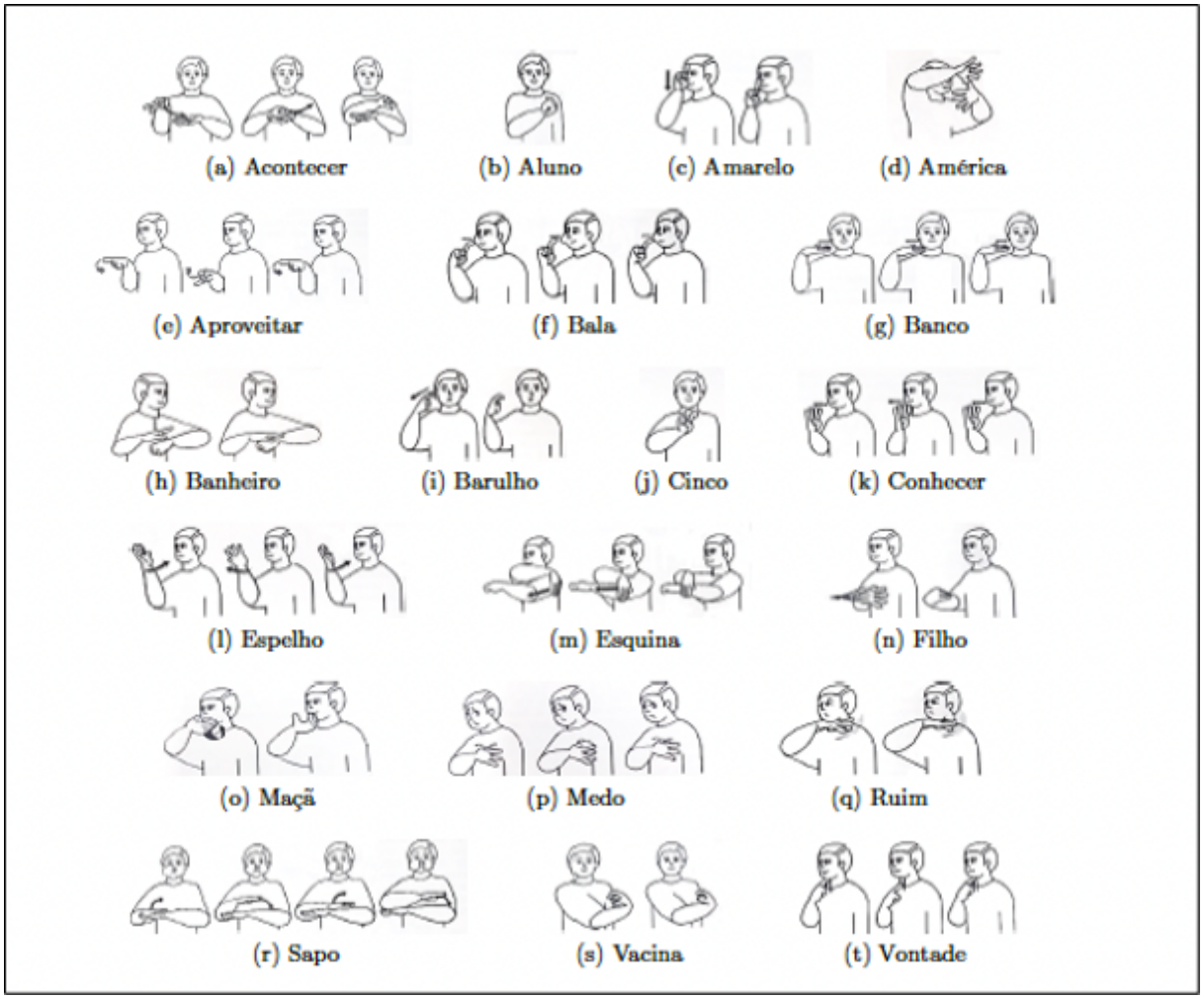

Figura 3 – Sinais disponibilizados na base MINDS-Libras. Fonte:(REZENDE, 2021)

3 Definição do Problema

Durante a fase de definição do problema, os objetivos específicos da pesquisa foram cuidadosamente identificados. A análise inicial começou com a formulação de uma hipótese através da indução, a partir da qual foi observado que a aplicação de técnicas computacionais poderia desempenhar um papel significativo na promoção da inclusão social de pessoas com deficiência auditiva e de fala (LAKATOS; MARCONI, 2003). O foco principal desta etapa foi avaliar as dimensões da inclusão social, com ênfase na Língua Brasileira de Sinais (LIBRAS), e explorar como as tecnologias emergentes poderiam facilitar uma maior integração dessas pessoas na sociedade.

Foi identificado que, embora a Língua Brasileira de Sinais seja um componente essencial para a inclusão, persistem diversas barreiras sociais que dificultam a comunicação. Esses obstáculos incluem não apenas a complexidade e o desafio associados ao aprendizado de um novo idioma, mas também a falta de interesse e incentivo para a aprendizagem por parte da sociedade em geral. Essas barreiras evidenciam a necessidade urgente de soluções que tornem a comunicação mais acessível e efetiva.

Para enfrentar esses desafios, a pesquisa adotou uma abordagem metodológica voltada para a classificação de sinais em LIBRAS, integrando técnicas avançadas de aprendizado de máquina. Essa escolha foi motivada pelo potencial dessas técnicas para melhorar a precisão e a eficácia na comunicação surdo-ouvinte, facilitando o processo de inclusão.

A metodologia empregada seguiu os princípios clássicos da ciência de dados voltada para problemas de classificação. O primeiro passo consistiu na seleção e identificação de uma base de dados adequada para a pesquisa. Após a definição da base de dados, foram exploradas metodologias específicas descritas na literatura acadêmica, com o objetivo de aplicar as melhores práticas para a classificação de vídeos contendo sinais em LIBRAS.

A etapa subsequente envolveu a escolha de uma base de dados apropriada e a revisão de trabalhos acadêmicos relacionados ao tema. Essa análise foi fundamental para a seleção do modelo de classificação mais adequado. O modelo escolhido foi então treinado utilizando a base de dados selecionada, e os resultados obtidos foram minuciosamente avaliados para determinar a eficácia da abordagem adotada.

Esse processo metodológico permitiu uma avaliação detalhada dos modelos de classificação, fornecendo *insights* valiosos sobre a eficácia das técnicas aplicadas e contribuindo para o avanço das soluções tecnológicas voltadas para a inclusão social de pessoas com deficiência auditiva e de fala.

Nesta etapa ainda foi realizada uma revisão da literatura, onde segundo NORONHA e FERREIRA (2000), a revisão da literatura é essencial para a elaboração de um texto científico, fornecendo uma perspectiva histórica sobre o tema e evidenciando novas ideias e métodos com maior ou menor evidência na literatura especializada. Neste sentido foi realizada a busca e análise de trabalhos com problemas semelhantes. Esta busca foi direcionada para identificar na literatura trabalhos envolvendo reconhecimento e classificação de vídeo com foco, especificamente na classificação de sinais em LIBRAS.

Neste processo foram analisados diversos trabalhos sobre o tema, onde podemos destacar o trabalho de Rezende (2021) que apresenta uma metodologia a qual foi aplicada com sucesso na classificação de sinais em LIBRAS, bem como disponibilizou uma base de dados para trabalhos com este tema.

4 Metodologia

Este capítulo apresenta as etapas utilizadas no processo de reconhecimento de língua de sinais. Este processo se baseia em uma metodologia geral de um processo envolvendo aprendizado de máquina e classificação de vídeos. Onde as seguintes etapas básicas foram seguidas: seleção da base de dados, preparação da base de dados, escolha e treinamento do modelo, avaliação dos modelos.

4.1 Seleção da base de dados

Neste trabalho, foi utilizado a base MINDS-Libras, composta por mais de 1100 amostras, com 20 sinais em LIBRAS gravados por doze pessoas diferentes usando sensores RGB-D e câmeras RGB. Os sinais foram selecionados por uma profissional proficiente em LIBRAS, com base na variabilidade dos parâmetros fonológicos, como configuração da mão, ponto de articulação, movimento das mãos, orientação da palma e expressões não-manuais (REZENDE, 2021).

A base inclui 115 GB de informações, abrangendo vídeos, dados dos sensores RGB-D, informações das articulações do corpo e pontos da face. Os dados são armazenados em vídeos no formato .mp4 e arquivos texto no formato.txt com dados das articulações e pontos da face. Para o presente trabalho, optou-se por utilizar apenas os dados dos arquivos de vídeo, por ser um formato mais comum e que não necessita de sensores adicionais para ser gerado, o que permite uma maior generalização.

A escolha da base MINDS-Libras se justifica pela sua adequação e relevância para os propósitos da pesquisa. Esta base é uma fonte rica e diversificada de sinais da Língua Brasileira de Sinais, proporcionando um ambiente de estudo e análise relevante para investigações voltadas para a comunicação com a comunidade surda. A qualidade e confiabilidade da base MINDS-Libras, assim como sua utilização em pesquisas anteriores, atestam sua relevância e validade para estudos relacionados à LIBRAS e comunicação visual gestual.

A base MINDS-Libras inclui documentação detalhada sobre o contexto de coleta dos dados, participantes envolvidos e procedimentos de gravação, o que facilita a manipulação e análise dos dados.

Embora a base MINDS-Libras possua informações de pontos de articulação e pontos da face, neste trabalho optamos em trabalhar apenas com os vídeos RGB-D em formato .mp4, pois este é o formato mais básico e para sua captura não há a necessidade de sensores adicionais.

4.2 Preparação da base de dados

A preparação da base de dados é uma etapa crucial em projetos de aprendizado de máquina, frequentemente considerada uma das mais difíceis e importantes. Esta etapa pode incluir definição do problema, preparação dos dados, validação dos modelos e definição do modelo final (BROWNLEE, 2020).

A preparação dos dados é essencial, pois modelos de aprendizado de máquina esperam dados numéricos e podem ter requisitos específicos, além de precisarem de tratamento de ruídos e erros estatísticos. As etapas incluem limpeza dos dados, seleção de variáveis e redução de dimensionalidade.

Em problemas de classificação de vídeos, é necessário converter vídeos em um conjunto de *frames*, dado que muitos *frames* têm pouca variação. Por exemplo, um vídeo gravado a 30 *frames* por segundo possui 30 imagens por segundo, com pouca variação entre os *frames*. Variações significativas podem indicar mudanças de contexto, cena ou movimento (ASIM *et al.*, 2018).

Analisar um vídeo de 5 segundos pode envolver a análise de até 150 imagens. A redução da dimensionalidade é uma etapa importante para diminuir o tamanho e a quantidade de imagens a serem analisadas, o que reduz o tempo de processamento e aumenta a acurácia do modelo (KANISHKA *et al.*, 2023).

Para essa tarefa, foi utilizada a biblioteca Katma, desenvolvida pela KeplerLab e disponibilizada para a comunidade *Python* sob licença MIT. A biblioteca possui um módulo para extração de *key frames* em vídeos, baseado em agrupamento K-means com histograma de imagem, diferenças absolutas no espaço de cor LUV, e filtragem de brilho e contraste dos quadros extraídos (KEPLERLAB, 2024).

A escolha da biblioteca Katma se deu devido à sua facilidade de uso e ao fato de que a atividade de extração de *frames* não é o foco principal do trabalho. No entanto, essa abordagem tem algumas limitações quanto à capacidade de configuração do algoritmo utilizado.

Para a divisão da base de dados em treinamento e teste, inicialmente optou-se por uma separação de 80% para treinamento e 20% para testes. A seleção dos dados foi feita de forma aleatória, mas posteriormente ajustou-se para selecionar vídeos de alguns sinalizadores para treinamento e apenas um para teste.

Para a extração das características, foram testados alguns modelos CNN pré-treinados na base de dados ImageNet, incluindo DenseNet121, Inception V3 e VGG16. Os *key frames* extraídos foram ajustados de acordo com o tamanho esperado por cada modelo, sendo 128x128 *pixels* para DenseNet121 e VGG16 e 224x224 pixels para Inception V3. Os modelos retornam, respectivamente, 1024 características para DenseNet121, 2048 para Inception V3 e 1000 para VGG16. Todos os modelos utilizaram imagens com três dimensões de cor RGB.

4.3 Escolha e treinamento do modelo

Nesta etapa, foi realizado o processo de treinamento dos modelos identificados na literatura. Foram realizados experimentos para determinar os parâmetros a serem utilizados no processo de reconhecimento de sinais em LIBRAS, utilizando as extrações de características usando os modelos pré-treinados.

Atualmente, a principal abordagem para reconhecimento automático e classificação de imagens é o uso de técnicas de *deep learning* ou aprendizado profundo, especialmente redes neurais convolucionais (CNN). Essa abordagem é amplamente utilizada na literatura para a classificação do alfabeto manual em LIBRAS. No entanto, para a classificação de sinais, considerando os parâmetros formativos da linguagem, é necessário utilizar uma sequência de imagens com lógica temporal (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

Para o processamento de imagens com sequência temporal, que é o caso da problemática envolvendo o reconhecimento de língua de sinais, as abordagens mais comuns são o CNN-3D, o CNN+RNN e o *Transformer*. Neste trabalho, foram avaliadas as abordagens CNN+RNN e *Transformer*.

A escolha destes modelos foi baseada em algumas considerações fundamentais. Primeiramente, a combinação de Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN) oferece uma integração eficiente das capacidades de extração de características espaciais e análise temporal. Já o modelo *Transformer* tem se destacado devido à sua eficiência em lidar com grandes volumes de dados e à sua capacidade de capturar relações de longo alcance entre elementos de dados por meio de mecanismos de atenção. O modelo CNN-3D não foi avaliado neste trabalho CNN-3D pois o trabalho proposto por (REZENDE, 2021) já aborda este modelo com a base MINDS-libras Dataset, o que permite uma comparação com os modelos aqui avaliados.

A avaliação dessas duas abordagens permite uma análise abrangente e comparativa entre metodologias tradicionais e modernas. A escolha de explorar tanto CNN+RNN quanto *Transformers* proporciona uma visão detalhada sobre o desempenho, a eficiência computacional e a robustez de cada abordagem, permitindo uma comparação eficaz das soluções para o problema em questão. Esta comparação é essencial para entender qual abordagem oferece os melhores resultados em termos de precisão, eficiência e capacidade de adaptação a diferentes tipos de dados e contextos.

4.4 Avaliação dos modelos

Na etapa de avaliação dos modelos, foram realizados testes para avaliar o desempenho dos modelos e determinar os parâmetros mais adequados para o reconhecimento de sinais em LIBRAS. Foram analisados os dados de entrada e saída dos algoritmos, comparadas as acurácias retornadas pelos modelos e realizados testes com diferentes quantidades de épocas (BACKES; JUNIOR, 2016).

Os resultados foram comparados entre diferentes modelos, analisando os *key frames* identificados para sinais com maior taxa de acerto e erro. A análise dos *key frames* visou identificar se o frame selecionado permite uma classificação correta do sinal.

Métodos estatísticos foram aplicados para analisar dados quantitativos, como taxas de acerto e tempo de resposta, e comparar os resultados de treinamento com os resultados de teste de outros trabalhos científicos. A aplicação de modelos estatísticos é crucial para identificar objetos e sinais em imagens naturais, lidando com grandes variações em imagens.

Após a análise estatística, foi possível avaliar o modelo que apresentou os melhores resultados com base nos parâmetros selecionados. O modelo de reconhecimento baseado na técnica escolhida foi implementado. Segundo RAVPREET e SINGH (2023), redes neurais convolucionais profundas têm demonstrado excelente performance em processamento de vídeo, segmentação de imagens e classificação, assim como em processamento de linguagem natural (RAVPREET; SINGH, 2023).

Finalmente, foram integrados os resultados quantitativos para uma interpretação abrangente sobre a utilidade prática da técnica. Nesta etapa, foram analisados os dados e resultados conforme os objetivos da pesquisa, com foco na comparação dos resultados e identificação das melhores soluções (WAZLAWICK, 2009).

5 Resultados

Este capítulo apresenta os resultados obtidos com base nos modelos propostos, utilizando a metodologia proposta no capítulo 3. Avaliamos também o resultado do tratamento dos dados e extração dos *key frames*, e como isso impacta no resultado do treinamento do modelo. Desta forma avaliamos os resultados quanto ao processo de sumarização de vídeo, quanto a extração das características dos *frames*, utilizando os modelos pré treinados, treinamento do modelo, validação e teste do modelo.

5.1 Sumarização de vídeo

Durante o processo de preparação da base de dados foi realizado o processo de extração dos *frames* de cada vídeo através da biblioteca *OpenCV (Open Computer Vision Library)*. Esta biblioteca *OpenCV* foi originalmente desenvolvida pela Intel, sendo multiplataforma, livre ao uso comercial e acadêmico, possuindo módulos de processamento de imagens e video I/O sendo frequentemente aplicada na área de visão computacional. (BRADSKI; PISAREVSKY; BOUGUET, 2006)

Foi realizada a leitura dos vídeos e cada frame foi transformado em um vetor de três dimensões de cores, *Red, Green e Blue*. Criando assim uma base de dados em quatro dimensões, onde a primeira dimensão corresponde a sequência de *frames*, e para cada quadro três dimensões de cores representando cada *pixel* da imagem.

Ainda durante a leitura dos vídeos foi realizado o redimensionamento das imagens. A base de dados disponibiliza os vídeos na resolução 1920 x 1080 *pixels*, e os modelos de extração de características utilizados trabalham com imagens em tamanho menor, foi necessário realizar este redimensionamento para o tamanho 224x224 que é o suportado pelos modelos utilizados.

De forma geral os sinais disponibilizados na base de dados, podem ser classificados com apenas 5 imagens em uma sequência temporal como apresentado na Figura 3. Para otimizar a etapa de extração de características e melhorar a acurácia do modelo é necessário realizar a extração dos quadros que representam o sinal a ser classificado. Sendo este o processo de sumarização de vídeo.

```
if __name__ == "__main__":

  # Inicializar módulo de vídeo
  vd = Video()

  # Número de frames
  no_of_frames_to_returned = 20

  #/content/drive/MyDrive/tcc/Sinalizador01/01Acontecer/1-01Acontecer_1RGB.mp4
  # Inicializar o  diskwriter para salvar os dados
  for index, x in enumerate(df['keyframes']):
    diskwriter = KeyFrameDiskWriter(location=x)
    video_file_path = df['video_name'][index]
    vd.extract_video_keyframes(
       no_of_frames=no_of_frames_to_returned, file_path=video_file_path,
       writer=diskwriter
    )
```

Figura 4 – Extração dos *key frames* usando a biblioteca Katna. Fonte: Autor

O processo de sumarização foi executado para cada video de cada sinalizador no formato MP4 disponibilizado na base de dados. Foi utilizada a biblioteca *Katna* com esta finalidade. Esta biblioteca possui um módulo de extração de *key frames*, o qual recebe como parâmetro o endereço do vídeo, o número máximo de *frames* que deseja que seja extraído e o caminho onde a biblioteca deve salvar os *key frames*. Conforme pode ser visualizado na Figura 4. Cada *frame* é salvo na pasta em formato JPEG. O processo de sumarização em geral levou 60 minutos para processar um conjunto de 100 vídeos, que é a quantidade de vídeo disponibilizada na base de dados para cada sinalizador.

A fim de validar a sumarização definimos inicialmente o parâmetro de número máximo de *frames* em 20. Porém em todos os testes a seleção efetiva foi de 3 a 5 *frames* por vídeo. O número máximo de *frames* é o único parâmetro fornecido pela biblioteca para ser configurado. Diante disso, foi realizada uma análise da biblioteca e ela utiliza o processo de clusterização através do algoritmo k-means para separação dos principais quadros, este processo ocorre após realizar tratativas de contraste, brilho e desfoque nas imagens. Ela utiliza o algoritmo k-means baseado no histograma de cores, selecionando os quadros que possuem maior variação de cores. Como as imagens em sua maioria tem pouca variação de cores entre os quadros, faz com que poucos quadros sejam selecionados. O parâmetro em questão não influencia na clusterização, sendo apenas um limitador da quantidade máxima de quadros após o processo de seleção.

Na Figura 5 podemos verificar os *key frames* extraídos para o sinal Aproveitar, em vídeos de sinalizadores diferentes.

Além disso antes de realizar a extração das características foi necessário proceder com a leitura de cada imagem extraída na respectiva pasta, e transformá-la em um array sequencial de imagens. Tal processo foi realizado para que seja possível aplicar as imagens em modelos de extração de características e no modelo de treinamento baseado em dados sequenciais.

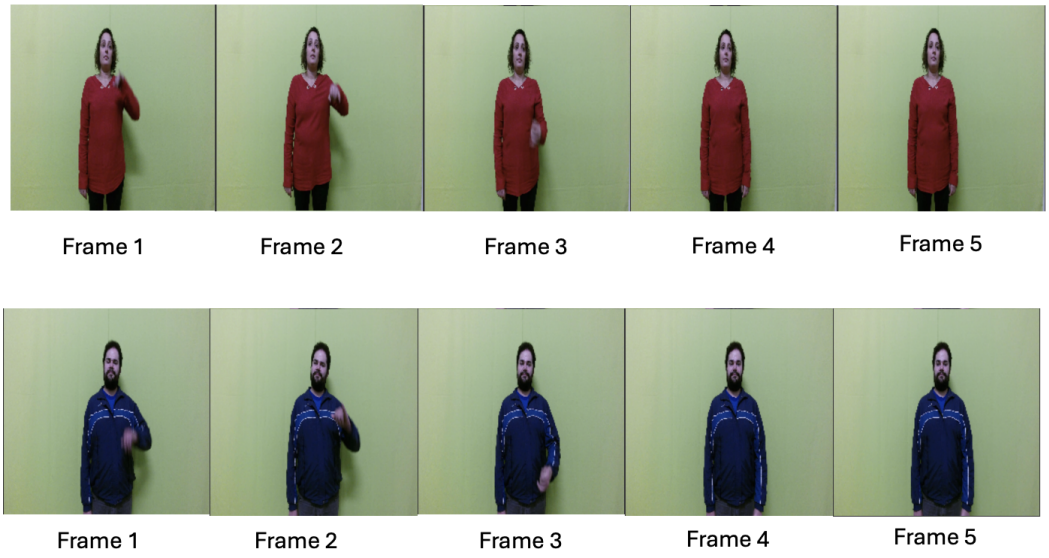

Figura 5 – Comparação dos *Key Frames* extraídos para o sinal Aproveitar nos sinalizadores 1 e 4

Posteriormente, a análise das amostras dos *frames* extraídos revelou que nem sempre a seleção realizada correspondia aos *frames* mais informativos para alguns sinais, influenciando na identificação precisa destes. Em particular, observou-se que os primeiros e últimos *frames* frequentemente representavam o sinalizador em uma posição neutra, com os braços abaixados. Esses *frames*, que marcam o início e o fim do sinal, não forneciam informações relevantes para a classificação.

Diante dessas limitações da biblioteca, para trabalhos futuros, a abordagem proposta por Freitas *et al.* (2014), baseada no Problema da Diversidade Máxima, pode ser utilizada. Esta abordagem ao ser aplicado na sumarização de vídeo se baseia na máxima distância temporal e a máxima diferença entre cores RGB entre eles. Essa abordagem tem o potencial de proporcionar uma captura mais fiel dos sinais e, consequentemente, aprimorar a acurácia dos modelos utilizados.

5.2 Extração de características

Antes do treinamento do modelo foi necessário fazer a extração das características de cada *frame*. Para esta etapa geramos a extração das características utilizando os modelos DenseNet121, Inception V3 e VGG16.

Este processo foi realizado com base em modelos pré treinados na base ImageNet. Foi construído um *notebook* do Google Colab específico para este fim. Nesta etapa, foi realizado o sequenciamento das imagens extraídas na etapa de sumarização utilizando a biblioteca openCV.

DenseNet 121

```
[28] def build_feature_extractor():
        feature_extractor = keras.applications.DenseNet121(weights="imagenet",
                                                            include_top=False, pooling="avg",
                                                            input_shape=(IMG_SIZE, IMG_SIZE, 3))
        preprocess_input = keras.applications.densenet.preprocess_input

        inputs = keras.Input((IMG_SIZE, IMG_SIZE, 3))
        preprocessed = preprocess_input(inputs)

        outputs = feature_extractor(preprocessed)
        return keras.Model(inputs, outputs, name="feature_extractor")

    feature_extractor = build_feature_extractor()
```

Figura 6 – Extração de características com DenseNet121

No processo de extração de características, uma variável importante é a quantidade de *features* extraídas por cada modelo. O modelo DenseNet121, por exemplo, utiliza um tamanho de *features* igual a 1024. Essa quantidade é determinada pela arquitetura densa da rede, que promove uma conexão eficiente entre as camadas e facilita a reutilização de informações. O DenseNet121, com suas 121 camadas, é projetado para representar características relevantes dos dados de forma eficiente, mantendo a complexidade computacional gerenciável.

Já o Inception V3 adota uma abordagem diferente, com uma quantidade de 2048 *features*. Esta escolha é possível devido à sua arquitetura de múltiplos caminhos, que permite a extração de características em várias escalas. Esse modelo é capaz de capturar uma gama mais ampla de detalhes visuais, oferecendo um bom equilíbrio entre a expressividade das características e a eficiência do processamento.

O VGG16, por sua vez, produz um vetor de 1000 *features* na camada final de extração. Esta configuração corresponde ao número de classes no conjunto de dados ImageNet, para o qual o modelo foi originalmente treinado. A escolha desse número de características permite que o VGG16 mantenha uma representação rica das características visuais, alinhando-se com a configuração padrão do modelo e evitando uma dimensionalidade excessiva.

Além disso, foi definido um número máximo de 5 *frames* para a extração de características, com base na análise da eficiência da sumarização dos *frames* e na capacidade do modelo de aprender de forma eficaz a partir dos dados. Esta escolha busca capturar uma quantidade representativa de informações dos vídeos, sem sobrecarregar o processo de extração. Nas Figuras 6, 7, e 8, pode-se verificar o código referente a cada modelo de extração de características.

Após a extração das características dos vídeos, os dados foram organizados em um formato estruturado, representado por um array com as dimensões [X, Y, Z]. Aqui, X denota o número total de vídeos na base, Y indica a quantidade máxima de *frames* por vídeo e Z corresponde ao número de características extraídas para cada *frame*. Com esses dados estruturados, o próximo passo foi a separação dos conjuntos de treinamento e teste para a avaliação dos modelos de reconhecimento.

Inception V3

```
def build_feature_extractor():
    feature_extractor = keras.applications.InceptionV3(
                                                        weights="imagenet",
                                                        include_top=False,
                                                        pooling="avg",
                                                        input_shape=(IMG_SIZE, IMG_SIZE, 3),
                                                    )

    preprocess_input = keras.applications.densenet.preprocess_input

    inputs = keras.Input((IMG_SIZE, IMG_SIZE, 3))
    preprocessed = preprocess_input(inputs)

    outputs = feature_extractor(preprocessed)
    return keras.Model(inputs, outputs, name="feature_extractor")

feature_extractor = build_feature_extractor()
```

Figura 7 – Extração de características com InceptionV3

VGG16

```
def build_feature_extractor():
    feature_extractor = keras.applications.VGG16( include_top=True,
                                                   weights="imagenet",
                                                   input_tensor=None,
                                                   input_shape=(IMG_SIZE, IMG_SIZE, 3),
                                                   pooling=None,
                                                   classes=1000,
                                                   classifier_activation="softmax",
                                                 )
    preprocess_input = keras.applications.densenet.preprocess_input

    inputs = keras.Input((IMG_SIZE, IMG_SIZE, 3))
    preprocessed = preprocess_input(inputs)

    outputs = feature_extractor(preprocessed)
    return keras.Model(inputs, outputs, name="feature_extractor")

feature_extractor = build_feature_extractor()
```

Figura 8 – Extração de características com VGG16

Para garantir uma divisão representativa e balanceada, foi adotada uma abordagem de amostragem aleatória estratificada. Essa técnica assegurou que a distribuição das classes nos conjuntos de treinamento e teste fosse proporcional àquela encontrada na base de dados original. Em geral, a divisão seguiu a proporção comum de 80% para treinamento e 20% para teste, adaptando-se conforme a quantidade total de dados disponível e as necessidades específicas do estudo.

Com a divisão estabelecida, as características extraídas foram alocadas de acordo com os conjuntos definidos, mantendo a integridade da estrutura [X, Y, Z] em ambos os casos. Essa organização possibilitou que o treinamento dos modelos fosse realizado exclusivamente com os dados de treinamento, enquanto a avaliação de desempenho fosse efetuada com os dados de teste.

Além disso, foi realizada uma verificação para garantir que ambos os conjuntos, treinamento e teste, fossem equilibrados e refletissem adequadamente as diversas classes presentes na base. Esse cuidado é essencial para evitar o sobre ajuste dos modelos e assegurar uma boa capacidade de generalização.

Durante o treinamento dos modelos, foram observadas variações nas taxas de acurácia, que diferiram conforme o modelo de extração utilizado. De maneira geral, o modelo baseado na DenseNet121 destacou-se por apresentar os melhores resultados, refletindo uma eficácia superior na captura das características relevantes dos sinais. Em contraste, o modelo VGG16 mostrou um desempenho inferior, independentemente das abordagens de treinamento aplicadas.

Essa metodologia de separação e análise dos dados permitiu uma avaliação detalhada do desempenho dos modelos e contribuiu para a identificação de abordagens mais eficazes no reconhecimento de sinais.

5.3 Treinamento do Modelo

No treinamento do modelo realizamos testes utilizando o modelo *Transformer* e o modelo CNN+RNN, para cada um deles testamos com a técnica de extração de características baseadas nos modelos CNN pré treinados DenseNet121, InceptionV3 e VGG16. Onde apresentaram os seguintes resultados de acurácia descritos na Tabela 1,

	DenseNet	**Inception V3**	**VGG16**
Transformers	67.5%	32.5%	5%
CNN+RNN	39%	28.25%	6%

Tabela 1 – Comparação do percentual de acurácia entre os modelos

Nos testes, o modelo *Transformer* apresentou o melhor resultado se comparado ao modelo CNN+RNN, com uma acurácia de 67,5%, enquanto a acurácia máxima do modelo CNN+RNN foi de 39%, ambos utilizando as características extraídas através do modelo DenseNet121 e configurados para 150 épocas de treinamento. As épocas referem-se ao número de vezes que o algoritmo de treinamento percorre todo o conjunto de dados de treinamento. Em cada época, o modelo ajusta seus pesos com base nos erros observados, visando minimizar a função de perda. Portanto, um maior número de épocas geralmente permite que o modelo aprenda mais sobre os dados, desde que não haja sobre ajuste.

Durante o treinamento, foi configurada a base de validação como 15% dos dados totais para ambos os modelos, o que ajuda a monitorar o desempenho do modelo em dados não vistos e a evitar o sobre ajuste.

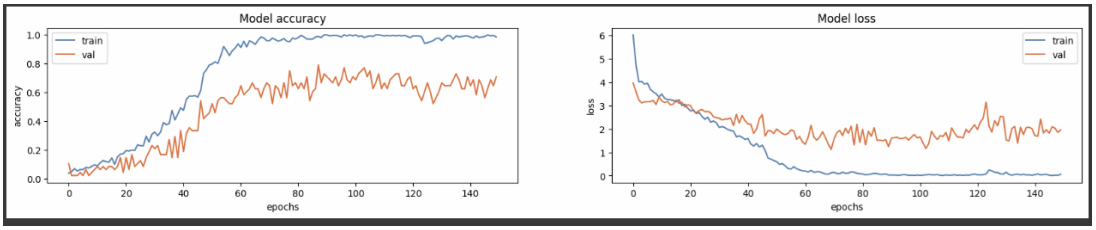

Figura 9 – Precisão e perda do modelo *Transformers* com DenseNet121

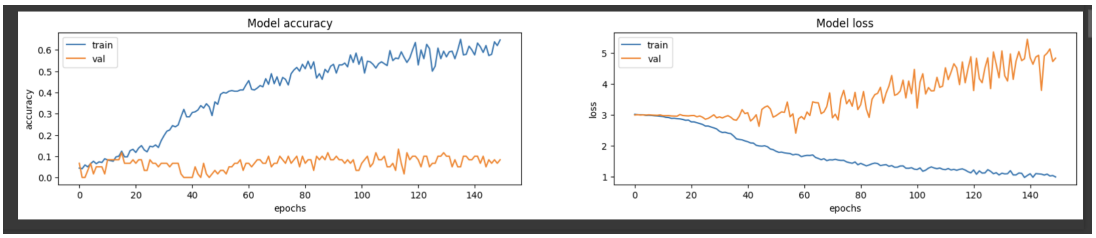

Figura 10 – Precisão e perda do modelo CNN+RNN com DenseNet121

Já o modelo VGG16 apresentou resultados muito abaixo dos demais, o que comprometeu uma avaliação adequada do modelo. Essa situação ocorreu para ambos os modelos de classificação *Transformer* e CNN+RNN, evidenciando que o VGG16, sob as condições estabelecidas, não conseguiu capturar as características que conseguisse classificar os sinais com boa acurácia.

Realizamos testes com diferentes valores de épocas, variando de 150 até 1500, para avaliar o desempenho dos modelos *Transformer* e CNN+RNN. Para o modelo *Transformer*, os resultados foram muito semelhantes em todas as variações de épocas, indicando que o número de épocas não teve um impacto significativo na acurácia do modelo. Já para o modelo CNN+RNN, observou-se um pequeno aumento na acurácia ao aumentar o número de épocas: a acurácia partiu de 32,5% com 150 épocas e alcançou 37,5% com 500 épocas. No entanto, acima de 500 épocas, até o máximo de 1500 épocas, a acurácia se estabilizou em 37,5%.

Esses resultados sugerem que, para os modelos testados, 150 épocas já seriam adequadas para alcançar uma performance satisfatória. Essa conclusão se aplica especificamente ao *Transformer* e ao CNN+RNN, ambos utilizando a extração de características baseada no DenseNet121 que foi o modelo que apresentou melhor acurácia.

5.4 Validação do Modelo

Nas Figuras 9, 10 utilizando tanto *Transformers* quanto CNN+RNN. Podemos verificar nestas figuras um gráfico comparativo entre a precisão*(accuracy)* e a perda*(loss)* do modelo utilizando *Transformers*, DenseNet121 e VGG para os conjuntos de treinamento e validação. Em ambos os cenários utilizado 150 épocas para o treinamento dos modelos. Como o conjunto de validação é balanceado em relação às classes, a precisão fornece uma representação imparcial do desempenho do modelo.

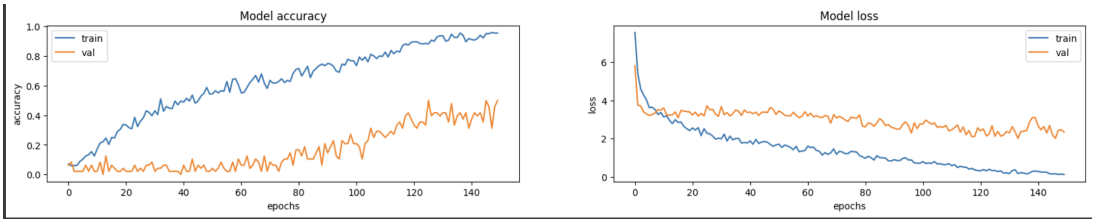

Figura 11 – Precisão e perda do modelo *Transformers* com InceptionV3

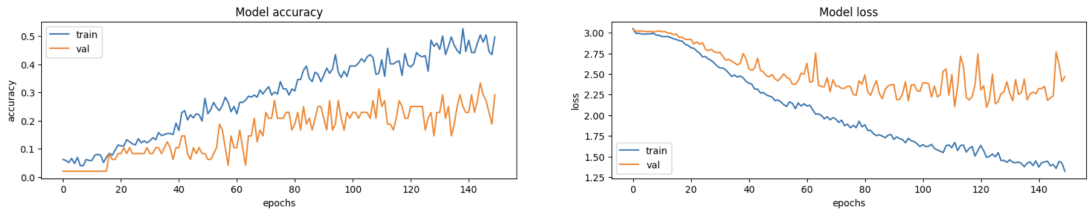

Figura 12 – Precisão e perda do modelo CNN+RNN com InceptionV3

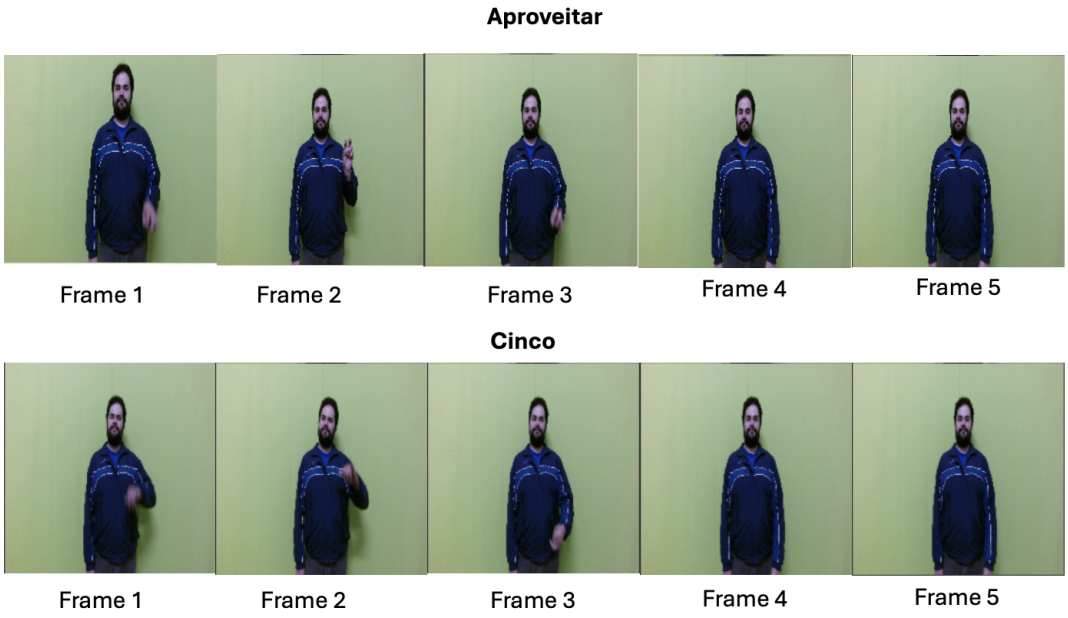

Figura 13 – Comparativo dos *frames* extraídos para o sinal Aproveitar e sinal Cinco

Ainda no processo de validação do modelo, ao analisar o modelo que apresentou melhor precisão e realizando testes com sinais identificamos que alguns sinais que apresentaram elevado percentual de acerto, ao ser utilizado por outro sinalizador apresentou erro na classificação. Isto aconteceu com os vídeos do sinal Ruim, que para o sinalizador 1 classificou erroneamente como Amarelo, para o sinalizador 4, classificou corretamente com 96% de certeza. O mesmo aconteceu com o sinal Aproveitar, que para o sinalizador 3 classificou com 100% de certeza, para o sinalizador 1 classificou como sinal Cinco com 59% de certeza. O que pode ser explicado principalmente pelo processo de seleção dos *key frames*, onde os *frames* selecionados entre os sinalizadores foram bem diferentes. Como pode-se perceber ao comparar os dois sinais com base nos *key frames* extraídos demonstrados na Figura 13.

6 Considerações Finais

Este estudo concluiu uma análise abrangente sobre a aplicação de técnicas de processamento de vídeo e aprendizado de máquina na identificação de sinais em LIBRAS. O foco principal foi desenvolver e avaliar um sistema capaz de reconhecer sinais em vídeos, com potencial aplicação em áreas como reconhecimento de linguagem de sinais e monitoramento de gestos.

O processo de sumarização de vídeo foi uma etapa fundamental. Utilizando ferramentas como OpenCV e a biblioteca Katna, conseguimos extrair *key frames* dos vídeos, permitindo uma representação compacta e informativa dos sinais. Apesar de algumas limitações na seleção automática de *frames,* como a não identificação sempre precisa dos momentos mais relevantes dos sinais, o processo geral mostrou-se eficiente na captura de aspectos essenciais dos vídeos. Porém para os vídeos de sinais em LIBRAS, nem sempre os *frames* escolhidos permitiam a diferenciação entre os diversos sinais. Em trabalhos futuros sugerimos avaliar outros métodos de extração dos *key frames*, de forma a avaliar o impacto desta etapa no resultado dos modelos.

A etapa subsequente envolveu a extração de características dos *frames* utilizando modelos pré-treinados como DenseNet121, Inception V3 e VGG16. Esses modelos, treinados em grandes conjuntos de dados como o ImageNet, mostraram-se capazes de capturar informações relevantes dos sinais presentes nos *frames*, fornecendo representações ricas e compactas para alimentar os modelos de aprendizado de máquina. Percebemos nesta etapa que o modelo VGG16 apresentou resultados muito baixos, o que pode estar diretamente ligado aos quadros selecionados na etapa anterior ou até mesmo por ser o modelo com o menor número de características selecionadas.

No treinamento dos modelos, foram exploradas duas abordagens principais: o modelo *Transformer* e o modelo CNN+RNN. Os resultados indicaram que o modelo *Transformer* teve um desempenho superior em comparação com o modelo CNN+RNN, alcançando uma acurácia máxima de 67,5%. Principalmente se usado em conjunto com o modelo DenseNet121 para extração das características. Este resultado ressalta a eficiência das arquiteturas de modelos baseadas em atenção para lidar com sequências de dados, como é o caso das sequências temporais de *frames* de vídeo.

A validação dos modelos também foi uma etapa crucial. A análise das métricas de precisão e perda nos conjuntos de treinamento e validação revelou que o desempenho do modelo *Transformers* em conjunto com o DenseNet121 foi consistente e que os modelos apresentaram uma capacidade satisfatória de generalização para dados não vistos, porém os demais modelos não apresentaram resultados satisfatórios, não alcançando o objetivo proposto, com alto grau de erro.

Em síntese, os resultados deste estudo indicam que a combinação de técnicas de sumarização de vídeo, extração de características e modelos de aprendizado de máquina oferece um potencial significativo para aplicações práticas na identificação de sinais em vídeos. O desenvolvimento desses sistemas tem implicações importantes, especialmente na acessibilidade e na comunicação para pessoas com deficiência auditiva, além de abrir caminho para aplicações em campos como interação humano-computador e monitoramento de gestos em ambientes inteligentes.

Como trabalhos futuros sugerimos algumas atividades como a revisão do processo de sumarização de vídeo utilizando outras abordagens, como através da abordagem do problema da diversidade máxima, buscando aumentar a acurácia dos modelos treinados, avaliar outras abordagens como CNN-3D em comparação com os modelos aqui analisados, aplicar os modelos treinamos em processo de classificação dos sinais utilizados na base de dados MINDS-Libras em tempo real.

Referências

ARAUJO, L. A. D. A proteção das pessoas com deficiência na constituição federal de 1988: A necessária implementação dos princípios constitucionais. In: **Volume V - Constituição de 1988 : O Brasil 20 anos depois. Os Cidadãos na Carta Cidadã**. Brasilia: Senado Federal, Instituto Legislativo Brasileiro, 2008. Disponível em: <https://www12.senado.leg.br/publicacoes/estudos-legislativos/tipos-de-estudos/outras-publicacoes/volume-v-constituicao-de-1988-o-brasil-20-anos-depois.-os-cidadaos-na-carta-cidada>.

ASIM, M.; ALMAADEED, N.; AL-MAADEED, S.; BOURIDANE, A.; BEGHDADI, A. A key frame based video summarization using color features. In: **2018 Colour and Visual Computing Symposium (CVCS)**. [S.l.: s.n.], 2018. p. 1–6.

ASSALEH, K.; AL-ROUSAN, M. Recognition of arabic sign language alphabet using polynomial classifiers. **EURASIP Journal on Applied Signal Processing**, v. 2005, n. 13, p. Z2136–2146, 2005.

BACKES, A. R.; JUNIOR, J. J. M. d. S. **Introdução à visão computacional usando Matlab**. São Paulo: Alta Books, 2016.

BATINI, C.; CAPPIELLO, C.; FRANCALANCI, C.; MAURINO, A. Methodologies for data quality assessment and improvement. **ACM Computing Surveys (CSUR)**, ACM, v. 41, n. 3, p. 1–52, 2009.

BENGIO, Y. Learning deep architectures for ai. **Computational Intelligence and Neuroscience**, Now Publishers, v. 2, n. 1, p. 1–127, 2019.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American**, v. 284, n. 5, p. 34–43, 2001.

BOJARSKI, M.; TESTA, D. D.; DWORAKOWSKI, D.; FIRNER, B.; FLEPP, B.; GOYAL, P.; JACKEL, L. D.; MONFORT, M.; MULLER, U.; ZHANG, J. *et al.* End to end learning for self-driving cars. **arXiv preprint arXiv:1604.07316**, 2016.

BORTHAKUR, D.; GRAY, J.; SARMA, J. S.; MUTHUKKARUPPAN, K.; SPIEGELBERG, N.; KUANG, H.; AIYER, A. Apache hadoop goes realtime at facebook. In: **Proceedings of the 2011 ACM SIGMOD International Conference on Management of data**. [S.l.: s.n.], 2007. p. 1071–1080.

BRADSKI, G. R.; PISAREVSKY, V.; BOUGUET, J. **Learning OpenCV: Computer Vision with the OpenCV Library**. [S.l.]: Springer, 2006.

BRASIL. **Decreto nº 5.626, de 22 de dezembro de 2005**. 2005. Regulamenta a Lei nº 10.436, de 24 de abril de 2002, que dispõe sobre a Língua Brasileira de Sinais - Libras. Brasília, DF.

BROWNLEE, J. **Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python**. [S.l.]: Machine Learning Mastery Pty. Ltd., 2020. ISBN 978-1-9993-4889-9.

CAPOVILLA, F. C.; MARTINS, A. C.; OLIVEIRA, W. G. S. Criando dicionários de línguas de sinais: modelos iconográfico, linguístico e contemporâneo. **International Journal of Operations & Production Management**, v. 18, n. 2, p. 152–169, 2017. ISSN 1809-4139. Disponível em: <https://doi.org/10.5935/cadernosdisturbios.v18n2p152-169>. Acesso em: 28 dez. 2023.

CARREIRA, J.; ZISSERMAN, A. Carreira et al. (2017). In: IEEE. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.], 2017. p. 2625–2634.

CARUANA, R.; GEHRKE, J.; KOCH, P. *et al.* Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: **Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2015. p. 1721–1730.

CHEN, C. H.; YANG, H.; HSU, Y. S. Reconhecimento de língua de sinais baseado em sensores wearables usando modelo de markov oculto. **Journal of Ambient Intelligence and Humanized Computing**, Springer, v. 8, n. 3, p. 319–328, 2017. Disponível em: <https://ieeexplore.ieee.org/document/7910696>.

CHO, K.; MERRIëNBOER, B. van; GULCEHRE, C.; BOUGARES, F.; SCHWENK, H.; BAHDANAU, D.; BENGIO, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. **Université de Montréal, Université du Maine, Jacobs University**, 2014.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. Imagenet: A large-scale hierarchical image database. **2009 IEEE conference on computer vision and pattern recognition**, p. 248–255, 2009.

DONAHUE, J.; HENDRICKS, L. A.; ROHRBACH, M.; VENUGOPALAN, S.; GUADARRAMA, S.; SAENKO, K.; DARRELL, T. **Long-term Recurrent Convolutional Networks for Visual Recognition and Description**. 2016. Disponível em: <https://arxiv.org/abs/1411.4389>.

DUARTE, S. B. R. e. a. **Aspectos históricos e socioculturais da população surda**. Rio de Janeiro: [s.n.], 2013. 1713-1734 p.

ELMASRI, R.; NAVATHE, S. B. **Fundamentals of database systems**. [S.l.]: Pearson, 2015.

ESCALERA; SERGIO; VASSILIS ATHITSOS E GUYON, I. Challenges in multi-modal gesture recognition. in gesture recognition. **Journal of Machine Learning Reserach**, Journal of Machine Learning Research, Springer, 2016.

ESTEVA, A.; KUPREL, B.; NOVOA, R. A.; KO, J.; SWETTER, S. M.; BLAU, H. M.; THRUN, S. Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, v. 542, n. 7639, p. 115–118, 2017.

FERRAIOLO, D.; CUGINI, J.; KUHN, R. Role-based access control (rbac): Features and motivations. In: **Proceedings of the 11th annual computer security applications conference**. [S.l.: s.n.], 2001. p. 241–248.

FERREIRA-BRITO, M. **Inclusão e Acessibilidade dos Surdos na Sociedade**. [S.l.]: Editora ABC, 1993.

FERREIRA, J. d. J.; BARBOSA, P. A.; MARTINO, J. M. D.; WILL, A. D.; OLIVEIRA, M. R. N. d. S.; SILVA, I. R.; XAVIER, A. N. Análise do papel das expressões não manuais na intensificação em libras. **DELTA: Documentação de Estudos em Linguística Teórica e Aplicada**, v. 34, n. 3, p. 723–745, 2018. Disponível em: <https://www.scielo.br/j/delta/a/ZkwFT3Nh3ncD4z8TpzjDK4d/>.

FREITAS, A. R. R.; GUIMARãES, F. G.; SILVA, R. C. P.; SOUZA, M. J. F. Memetic self-adaptive evolution strategies applied to the maximum diversity problem. **Optimization Letters**, v. 8, n. 2, p. 705–714, 2014. 2014a.

GARCIA-MOLINA, H.; ULLMAN, J. D.; WIDOM, J. **Database systems: The complete book**. [S.l.]: Pearson Education, 2002.

GESSER, Audrei. **LIBRAS? que língua é essa?: Crenças e preconceitos em torno da língua de sinais e da realidade surda.** São Paulo: Parábola Editorial, 2009.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA: MIT Press, 2016. ISBN 978-0262035613.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: Concepts and techniques**. [S.l.]: Elsevier, 2011.

HASAN, M. and MISRA, M. Hand gesture modeling and recognition using geometric features: A review. **Canadian Journal on Image Processing and Computer Vision**, n. 3, p. 12–26, 2012.

HAZARD, D.; FILHO, T. A. G.; REZENDE, A. L. A. **Inclusão digital e social de pessoas com deficiência**. 3. ed. Brasilia: UNESCO, 2007. 73 p. ISBN 370.6813. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000160012.locale=en>. Acesso em: 20 dez. 2023.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. **Proceedings of the IEEE conference on computer vision and pattern recognition**, p. 770–778, 2016.

HESTNESS, J.; NARANG, S.; ARDALANI, N.; DIAMOS, G.; JUN, H.; KIANINEJAD, H.; PATWARY, M. M. A.; YANG, Y.; ZHOU, Y. **Deep Learning Scaling is Predictable, Empirically**. 2017. Disponível em: <https://arxiv.org/abs/1712.00409>.

HOLSTEIN, K.; VAUGHAN, J. W.; III, H. D. Improving fairness in machine learning systems: What do industry practitioners need? **CHI Conference on Human Factors in Computing System**, n. 600, p. 1–16, 2019.

HUANG, G.; LIU, Z.; MAATEN, L. van der; WEINBERGER, K. Q. Densely connected convolutional networks. **Proceedings of the IEEE conference on computer vision and pattern recognition**, p. 4700–4708, 2017.

INFANTINO, I.; RIZZO, R.; GAGLIO, S. A framework for sign language sentence recognition by commonsense context. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v. 37, n. 5, p. 1034–1039, 2007.

JONES, A.; PATEL, S. Sign languages and their universality. **Journal of Linguistic Studies**, v. 8, n. 2, p. 45–62, 2015.

KANISHKA, G.; REDDY, K. J.; NOWREEN, A.; SATAKARNI, D.; JAGADEESH, M. V. S. Video summarization using keyframe analysis. **International Journal of Engineering Technology and Management Sciences**, v. 7, n. Special Issue 1, p. 521–524, April 2023. ISSN 2581-4621. Impact Factor Value: 5.672, Page 521-524.

KARPATHY, A.; FEI-FEI, L. Large-scale video classification with convolutional neural networks. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, p. 1725–1732, 2014.

KEPLERLAB. **Katna documentation**. 2024. Disponível em: <https://katna.readthedocs.io/en/latest/index.html>.

KHAN, A.; AWAIS, M.; QURESHI, A. S. Challenges and solutions for dataset standardization in machine learning applications. **Journal of Machine Learning Research**, Microtome Publishing, v. 23, n. 45, p. 678–693, 2022.

KIM, B. B.; CHOI, S.; LEE, H. Reconhecimento em tempo real da língua de sinais usando kinect. **IEEE Transactions on Multimedia**, IEEE, v. 21, n. 5, p. 1234–1245, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8706368>.

KOLLER, O.; NEY, H.; BOWDEN, R. Deep sign language recognition: Understanding the role of lstm in recognizing signs. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. IEEE, 2016. p. 359–368. Disponível em: <https://ieeexplore.ieee.org/document/7532511>.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. **Imagenet Classification with Deep Convolutional Neural Networks**. [S.l.]: MIT Press, 2012. 1097–1105 p.

KWON, H.; KIM, B.; CHO, Y. 3d convolutional networks for sign language recognition. **IEEE Transactions on Multimedia**, IEEE, v. 21, n. 9, p. 2116–2126, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8685783>.

LAKATOS, E. M.; MARCONI, M. D. A. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas, 2003.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, IEEE, v. 86, n. 11, p. 2278–2324, 1998.

LI, W.; LIU, C.; XU, Y. Skeleton-based sign language recognition with graph convolutional networks. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 29, n. 10, p. 4412–4423, 2018. Disponível em: <https://ieeexplore.ieee.org/document/8340641>.

LI, X.; WU, X.; ZHANG, Y. Deep learning for sign language recognition: A review. **IEEE Transactions on Neural Networks and Learning Systems**, v. 31, n. 8, p. 2612–2630, 2020.

LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; HAYS, J.; PERONA, P.; RAMANAN, D.; DOLL'aR, P.; ZITNICK, C. L. Microsoft coco: Common objects in context. **European conference on computer vision**, Springer, p. 740–755, 2014.

LITJENS, G.; KOOI, T.; BEJNORDI, B. E.; SETIO, A.; G., J. A. A. W. M. J. A.; T., M. L. M.; A., T. D.; S., F. A.; P., M.; L., L.; T., L. A survey on deep learning in medical image analysis. **Medical Image Analysis**, Elsevier, v. 42, p. 60–88, 2017.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. **Proceedings of the IEEE conference on computer vision and pattern recognition**, p. 3431–3440, 2015.

MAHAJAN, A.; GIRSHICK, R.; RAMANATHAN, V.; HE, K.; PALURI, M.; LI, Y.; BHARAMBE, A.; MAATEN, L. van der. Mahajan et al. (2018). In: **Proceedings of the European Conference on Computer Vision (ECCV)**. [S.l.: s.n.], 2018. p. 382–398.

MCILRAITH, S. A.; SON, T. C.; ZENG, H. Semantic web services. **IEEE Intelligent Systems**, v. 16, n. 2, p. 46–53, 2001.

MNIH, V.; KAVUKCUOGLU, K.; SILVER, D. *et al.* Human-level control through deep reinforcement learning. **Nature**, Nature Publishing Group, v. 518, n. 7540, p. 529–533, 2015.

NG, J. Y.; HAUSKNECHT, M.; VIJAYANARASIMHAN, S.; VINYALS, O.; MONGA, R.; TODERICI, G. Ng et al. (2015). **arXiv preprint arXiv:1505.05754**, 2015.

NGUYEN, T. P.; NGUYEN, H. T.; PHAM, D. T. Reconhecimento de língua de sinais usando técnicas de aprendizado profundo. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 42, n. 8, p. 2014–2025, 2020. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231220302148>.

NORONHA, D. P.; FERREIRA, S. M. S. P. **Revisões de literatura**. 5. ed. Belo Horizonte: UFMG, 2000.

OPENCV. Opencv. 2024. Disponível em: <https://opencv.org/about/>. Acesso em: 20 Maio. 2024.

PAN, S. J.; YANG, Q. A survey on transfer learning. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 22, n. 10, p. 1345–1359, 2010.

RAO, G.; KISHORE, P. V. V. Selfie video based continuous indian sign language recognition system. **Ain Shams Engineering Journal**, 2017. Disponível em: <https://api.semanticscholar.org/CorpusID:57735495>.

RAVPREET, K.; SINGH, S. A comprehensive review of object detection with deep learning. Digital Signal Processing, 2023. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S1051200422004298>. Acesso em: 17 dez. 2023.

REDDY, C. K.; AGGARWAL, C. C.; REDDY, V. S. **Data quality: Concepts, methodologies and techniques**. [S.l.]: Springer, 2018.

REZENDE, T. M. **Reconhecimento Automático de Sinais da Libras: Desenvolvimento da Base de Dados MINDS-Libras e Modelos de Redes Convolucionais**. Belo Horizonte: Universidade Federal de Minas Gerais, 2021. Disponível em: <https://repositorio.ufmg.br/handle/1843/39785>. Acesso em: 05 maio. 2024.

REZENDE, T. M.; SILVA, G. M. A.; GUIMARãES, F. G. Development and validation of a brazilian sign language database for human gesture recognition. Neural Comput Applic, v. 33, n. 39, p. 10449–10467, 2021. Disponível em: <https://doi.org/10.1007/s00521-021-05802-4>. Acesso em: 03 jan. 2024.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Why should i trust you?" explaining the predictions of any classifier. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, ACM, p. 1135–1144, 2016.

ROCHA, L. R. M. da; PASIAN, M. S. A proteção das pessoas com deficiência na constituição federal de 1988: A necessária implementação dos princípios constitucionais. Educação em Revista, n. 39, 2023. Disponível em: <http://dx.doi.org/10.1590/0102-469840565>. Acesso em: 28 dez. 2023.

RODRIGUES, M. **Inclusão e Acessibilidade dos Surdos na Sociedade**. 2008. São Paulo: Editora ABC.

SAMEK, W.; MONTAVON, G.; VEDALDI, A.; HANSEN, L. K.; MÜLLER, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. **arXiv preprint arXiv:1708.08296**, 2017.

SANTOS, F. M.; COSTA, A. M.; SILVA, R. T. Reconhecimento de libras usando dados de profundidade e cor. **Journal of Computer Vision**, Springer, v. 12, n. 4, p. 455–466, 2018. Disponível em: <https://link.springer.com/article/10.1007/s10559-018-9891-0>.

SARMA, N.; ZHANG, L.; LIU, S. Deep sign language recognition using rnns. **International Journal of Computer Vision**, Springer, v. 128, n. 3, p. 457–468, 2020. Disponível em: <https://link.springer.com/article/10.1007/s11263-020-01373-1>.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Database system concepts**. [S.l.]: McGraw-Hill, 2010.

SILVA, A. M.; ALMEIDA, A. M. C.; CAMPOS, E. A. de; NEVES, I. J. C. Redes de apoio e pessoas com deficiência física: inserção social e acesso aos serviços de saúde. **Ciência & Saúde Coletiva**, v. 20, n. 1, p. 97–108, 2015. Disponível em: <https://www.scielo.br/j/csc/a/XkBwC5pGcwQfnzcbc3fhV3h/?lang=pt>. Acesso em: 07 julho 2024.

SILVA, E. Evolução da língua brasileira de sinais: Um estudo de caso. **Revista de Linguística Aplicada**, v. 20, n. 3, p. 87–102, 2018.

Simons, G. and Fennig, C. **Simons e Fennig (2018)**. [S.l.]: Editora QRS, 2018.

SIMONYAN, K.; ZISSERMAN, A. Simonyan et al. (2014). **arXiv preprint arXiv:1409.1556**, 2014.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

SMITH, J. Communication and human senses. **Journal of Communication Studies**, v. 15, n. 1, p. 32–45, 2010.

SOUZA TANYA A. FELIPE DE; MONTEIRO, M. S. **Libras em Contexto: Curso Básico - Livro do Professor**. 6. ed. Brasilia: MEC, 2007. 448 p. ISBN 376.33.

STARNER, T.; WEAVER, J.; PENTLAND, A. A wearable computer based american sign language recognizer. In: **Digest of Papers. First International Symposium on Wearable Computers**. [S.l.: s.n.], 1997. p. 130–137.

STELLE, T. G.; STRIEICHEN, E. M. Os principais mitos sobre os surdos e a língua de sinais. **EDUCERE - XI Congresso Nacional de Educação**, XI Congresso Nacional de educação, 2013. Disponível em: <https://www.libras.com.br/download-files/libras/os-principais-mitos-sobre-os-surdos-e-a-lingua-de-sinais.pdf>. Acesso em: 5 maio. 2024.

STOKOE, J. **Communication and Human Senses**. [S.l.]: Journal of Communication Studies, 1960.

STONEBURNER, G.; GOGUEN, A.; FERINGA, A. **NIST SP 800-53: Recommended security controls for federal information systems**. [S.l.], 2002.

SZEGEDY, C.; IOFFE, S.; VANHOUCKE, V. Rethinking the inception architecture for computer vision. **Proceedings of the IEEE conference on computer vision and pattern recognition**, p. 2818–2826, 2016.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. Going deeper with convolutions. **Proceedings of the IEEE conference on computer vision and pattern recognition**, p. 1–9, 2015.

TRAN, D.; WANG, H.; TORRESANI, L.; RAY, J.; LECUN, Y.; PALURI, M. Tran et al. (2018). **arXiv preprint arXiv:1808.10393**, 2018.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, ; POLOSUKHIN, I. Attention is all you need. **Google Brain, Google Research, University of Toronto**, 2017.

VENUGOPALAN, S.; XU, H.; DONAHUE, J.; ROHRBACH, M.; MOONEY, R.; SAENKO, K. Translating videos to natural language using deep recurrent neural networks. In: MIHALCEA, R.; CHAI, J.; SARKAR, A. (Ed.). **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 1494–1504. Disponível em: <https://aclanthology.org/N15-1173>.

VIEVILLE, T.; CRAHAY, J. Using an hebbian learning rule for multi-class svm classifiers. **J Comput Neurosci**, v. 17, p. 271–289, 2004.

VONDRICK, C.; PIRSIAVASH, H.; TORRALBA, A. Vondrick et al. (2016). In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2016. p. 77–85.

VOULODIMOS, A.; DOULAMIS, N.; DOULAMIS, A.; PROTOPAPADAKIS, E. Deep learning for computer vision: A brief review. **Computational Intelligence and Neuroscience**, v. 2018, n. 1, p. 7068349, 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2018/7068349>.

WAZLAWICK, R. S. **Metodologia de pesquisa para ciência da computação**. Rio de Janeiro: Elsevier, 2009. 124 p. 6ª impressão. ISBN 9788535235227.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2016.

XU, Y.; ZHANG, M.; CHEN, L. Sign language recognition with transformer networks. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 32, n. 4, p. 1487–1496, 2021. Disponível em: <https://ieeexplore.ieee.org/document/9278523>.

YAN, S.; XIONG, Y.; LIN, D. Spatial-temporal graph convolutional networks for action recognition. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. IEEE, 2018. p. 3794–3803. Disponível em: <https://openaccess.thecvf.com/content_cvpr_2018/html/Yan_Spatial-Temporal_Graph_Convolutional_Networks_CVPR_2018_paper.html>.

YOSINSKI, J.; CLUNE, J.; BENGIO, Y.; LIPSON, H. How transferable are features in deep neural networks? **Advances in neural information processing systems**, v. 27, p. 3320–3328, 2014.

ZHOU, B.; ANDONIAN, A.; OLIVA, A.; TORRALBA, A. **Temporal Relational Reasoning in Videos**. 2018. Disponível em: <https://arxiv.org/abs/1711.08496>.

ZHU, W.; LAN, C.; XING, J.; ZENG, W.; LI, Y.; SHEN, L.; XIE, X. **Co-occurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks**. 2016. Disponível em: <https://arxiv.org/abs/1603.07772>.

ZHU, Y.; ZHANG, X.; LIU, J. Transfer learning for sign language recognition: A comprehensive survey. **Pattern Recognition**, Elsevier, v. 87, p. 1–15, 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320319300467>.