Uso de técnicas de Deep learning para o reconhecimento de sinais

André Luiz Santos Moreira da Silva[1], **Daniel Rocha Franca**[1], **Ligia Ferreira de Carvalho Gonçalves**[1]

[1]Instituto de Ciências Exatas e Informática – Pontifícia Universidade Católica de Minas Gerais (PUC Minas) – 30140-100 – Belo Horizonte – MG – Brazil

Abstract.

Resumo.

1. Contextualização

A comunicação para o ser humano é um dos principais pilares que sustentam seu estilo de vida, pois antes de ser racional nossa espécie é marcada pela característica da sociabilidade, e o principal método que temos é a fala. No entanto, a comunicação não se limita a fala, pode ser expressa de várias formas e para uma parcela significativa da população, essa forma são o uso de sinais.

A língua de sinais, é uma esfera universal em que cada país possui sua própria variação, a Organização Mundial de Surdos , estima que dentre os 70 milhões de surdos, coletivamente existam mais de 300 línguas de sinais distintos. Como exemplo podemos citar a ANL (American Sign Language), LIBRAS (Língua Brasileira de Sinais), dentre outras. De acordo com (Rastgoo et al. 2021), a interpretação dos sinais depende de cinco elementos: o formato das mãos, as expressões faciais, a orientação da palma, movimentações, a localização e as expressões faciais. Sistemas de reconhecimento de sinais não são uma linha de pesquisa recente, em 1986 (Zimmerman et al. 1986) propuseram uma interface que capturava informações em tempo real sobre gesticulações utilizando as mãos. Apesar dos recentes avanços, esses sistemas ainda apresentam limitações devido a forte influência de variáveis externas sobre seu desempenho, como iluminação, posicionamento da máquina, e até mesmo como o indivíduo construiu determinado sinal, a principal sendo a capacidade de traduzir frases completas em tempo real.

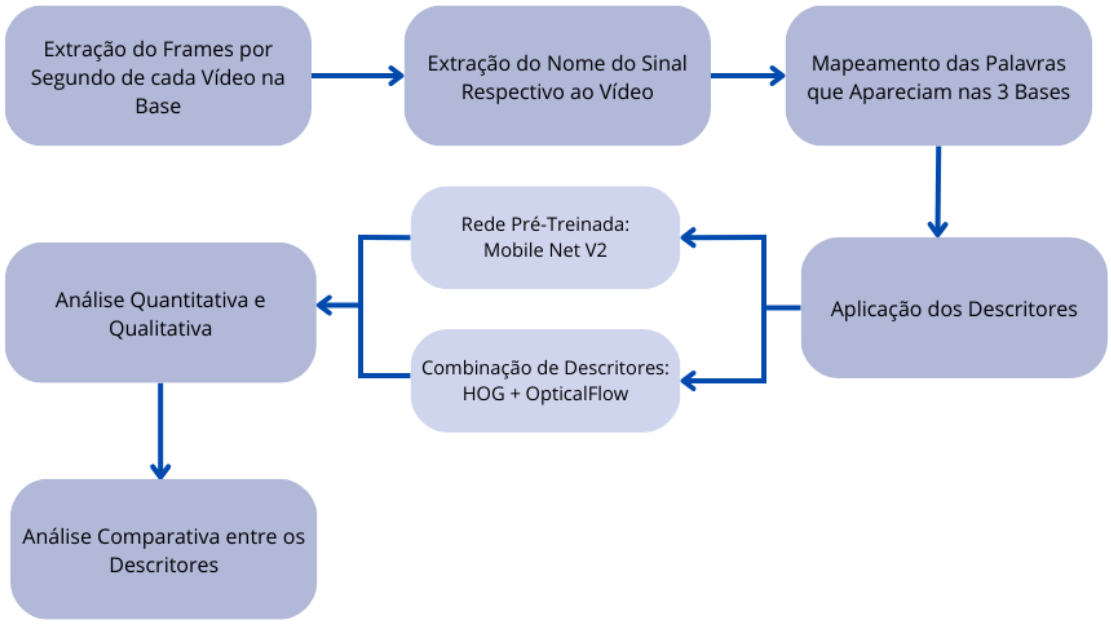
Desse modo a viabilização de métodos e processos responsáveis pelo reconhecimento desses sinais é fundamental para a manutenção do diálogo entre indivíduos. Suas aplicações ultrapassam o contexto da comunicação, estendendo-se a diversas áreas. Por exemplo, na educação, sistemas de reconhecimento de sinais podem facilitar o aprendizado de pessoas surdas, promovendo a inclusão em ambientes educacionais. Em serviços públicos, essas tecnologias garantem que informações e serviços governamentais sejam acessíveis a todos os cidadãos, independentemente de suas habilidades auditivas. Na saúde, a comunicação entre profissionais de saúde e pacientes surdos é aprimorada, assegurando um atendimento mais eficaz e humanizado. Além disso, no mercado de trabalho, ampliam-se as oportunidades de emprego para pessoas surdas, ao tornar ambientes corporativos mais inclusivos e adaptados às suas necessidades. Conforme destacado no estudo (Chaveiro et al. 2014), a surdez tem um impacto negativo sobre a qualidade de vida relacionada à saúde (QVRS) de pessoas surdas, sendo que sintomas de ansiedade

e depressão são mais acentuados nesse grupo e podem estar relacionados a dificuldades de comunicação. Portanto, investir em tecnologias e processos que aprimorem o reconhecimento e a interpretação das línguas de sinais é essencial para promover a inclusão social e garantir que indivíduos surdos possam participar plenamente de todas as esferas da sociedade.

2. Estudo e Comparação de Embeddings

O dataset separado para a realização deste trabalho, que está disponível em (https://libras.cin.ufpe.br/), é estruturado da seguinte maneira: três articuladores diferentes possuem um vídeo associado a cada um, representando o sinal de uma palavra. Ao todo, a base apresenta 1364 palavras, totalizando, com três bases, 4089 ocorrências. No entanto, durante os primeiros trabalhos com o dataset, algumas inconsistências foram descobertas; por exemplo, algumas palavras não estavam presentes em todas as bases. Essas questões precisaram ser tratadas.

Figure 1. Esquema Descritivo das Etapas do Processo de Ensemble



Desse modo, tendo como base a Figura 1, é possível abordar cada etapa de forma detalhada:

Extração dos Frames por Segundo de Cada Vídeo na Base: Para essa etapa, foi utilizada a biblioteca OpenCV, cuja principal característica é a leitura e exibição de imagens e vídeos, possibilitando, por isso, a captura e manipulação tanto em tempo real quanto de arquivos gravados. Dessa forma, foi usada para capturar um frame por segundo de cada vídeo. Após isso, foi criada uma subpasta para armazenar os frames de cada vídeo. O código pode ser encontrado em: Notebook de Extração de Frames

Extração do Nome do Sinal Respectivo ao Vídeo: Após a extração dos vídeos, foi usada a biblioteca Tesseract, que é uma ferramenta de reconhecimento óptico de caracteres (OCR), ou seja, uma API capaz de reconhecer caracteres a partir de arquivos de

imagem. Ela foi utilizada para remover os textos contidos em cada vídeo, permitindo, assim, que cada 'classe' fosse nomeada. O código pode ser encontrado em: Notebook Extração Nome Sinal

2.1. Implementação de Abordagens

A abordagem de combinar descritores para a extração de características não é novidade. No trabalho de (Konečný and Hagara 2013), os autores utilizaram a combinação dos descritores HOG (Histogram of Oriented Gradients) e HOF (Histogram of Optical Flow) para desenvolver um método de reconhecimento de gestos baseado em aprendizado a partir de um único exemplo (one-shot learning). Já em (Mota et al. 2014), os autores propuseram a concatenação de tensores individuais calculados com Optical Flow Approximation e HOG3D, formando um descritor global final utilizado para o reconhecimento de ações humanas. Essa abordagem se mostrou eficiente, uma vez que apresentou boas taxas de reconhecimento baseadas exclusivamente em vídeo.

O uso de redes pré-treinadas para extração de características de imagem é uma prática consolidada, amplamente adotada na literatura científica. Em trabalhos mais recentes, como o de (Zhan 2019), os autores propõem a fusão entre uma rede CNN tradicional e módulos baseados em Transformers, com o objetivo de extrair simultaneamente características locais e globais das imagens. Os resultados experimentais demonstram que a estrutura de rede híbrida contribuiu significativamente para a melhoria da acurácia no reconhecimento de gestos. Dessa forma, é possível afirmar que os dois métodos adotados neste trabalho possuem fundamentação teórica consistente, o que justifica suas respectivas escolhas.

Combinação de Descritores:

Neste experimento, o objetivo é combinar descritores extraídos de três bases de vídeos em LIBRAS, organizadas em diretórios. Para isso, utilizamos a rede neural pré-treinada ResNet50, que é especializada na extração de características de imagens. O modelo ResNet50, ao ser aplicado, gera um vetor de características para cada imagem, capturando informações relevantes, como formas, texturas e padrões. A extração dos descritores é feita a partir de até três frames por sequência de vídeo, escolhendo as imagens de cada diretório e aplicando a rede para gerar os vetores representativos de cada uma.

A abordagem adotada para a extração dos descritores inclui o uso de uma função personalizada que percorre os diretórios das três bases de dados, carrega as imagens, aplica o pré-processamento necessário (como redimensionamento e normalização) e então extrai os descritores usando o modelo ResNet50. Cada imagem é processada individualmente, e seu descritor resultante é armazenado em uma lista. Essa abordagem permite que todas as imagens sejam tratadas de forma independente, garantindo que o descritor de cada uma seja calculado e armazenado corretamente para futura análise.

Após a extração dos descritores, todos os vetores são combinados em um único array, onde cada linha representa os descritores de uma imagem. Esse array combinado é então salvo em um arquivo CSV, o que facilita o uso posterior dos dados em outros processos de análise, como visualização com técnicas de redução de dimensionalidade ou treinamento de modelos de aprendizado de máquina. O processo de combinação de descritores permite a criação de uma representação global das amostras das três bases,

facilitando a análise comparativa entre elas e demonstrando a efetividade da ResNet50 na captura de características relevantes das imagens de LIBRAS.

Descritor Baseado em uma Rede Neural: Para realizar a tarefa da descrição dos frames extraídos dos vídeos, escolhemos a MobileNetV2, uma rede convolucional utilizada em aplicações para dispositivos móveis, sendo, portanto, leve e prática. A MobileNetV2 é construída para lidar com tarefas envolvendo o processamento de imagens, algo demonstrado em seu trabalho de origem (Sandler et al. 2018), onde é testada e treinada com tarefas de visão computacional. Neste trabalho, buscamos a versão da MobileNetV2 que a biblioteca TensorFlow fornece, que segue a estrutura original, porém nos permite usar os pesos da rede treinada na base ImageNet, amplamente usada na visão computacional. Como desejamos apenas a habilidade de descrever os frames com a MobileNetV2 treinada na ImageNet, removemos as camadas de classificação da rede treinada, obtendo como resultado uma saída 3D que descreve os mapas de ativações que a rede obteve para aquele frame específico. Em seguida, tratamos essa saída adicionando à rede uma camada de Global Average Pooling 2D, para transformá-la em um vetor. Após aplicarmos o descritor sobre cada frame de cada vídeo, combinamos os vetores resultantes em uma matriz que representa todos os frames do vídeo. Após concluirmos a extração da descrição de cada frame, realizamos uma avaliação da qualidade dos embeddings gerados. Para isso, utilizamos três abordagens complementares: cálculo da similaridade intra e interclasse, visualização com t-SNE e classificação com regressão logística. Nos próximos tópicos vamos explorar cada abordagem realizada.

2.1.1. Análise da Qualidade dos Embeddings Rede Neural

Na primeira abordagem, calculamos a similaridade média dos vetores entre vídeos da mesma classe (intra-classe) e entre vídeos de classes diferentes (inter-classe), utilizando a métrica de similaridade cosseno. O resultado obtido foi uma similaridade média intra-classe de 0.7936 e uma similaridade média inter-classe de 0.8079, indicando uma proximidade elevada entre vídeos, independentemente da classe, o que pode sugerir que os embeddings não estão separando bem as categorias semânticas.

2.1.2. Visualização Rede Neural

Já na intenção de visualizar, aplicamos o algoritmo t-SNE para reduzir os vetores a duas dimensões e visualizar os agrupamentos no espaço vetorial. A visualização gerada (Figura **??**) mostra alguns agrupamentos visuais aparentes, mas também uma considerável sobreposição entre diferentes classes, o que reforça a hipótese de que o descritor ainda não está totalmente conseguindo capturar bem as diferenças entre as categorias.

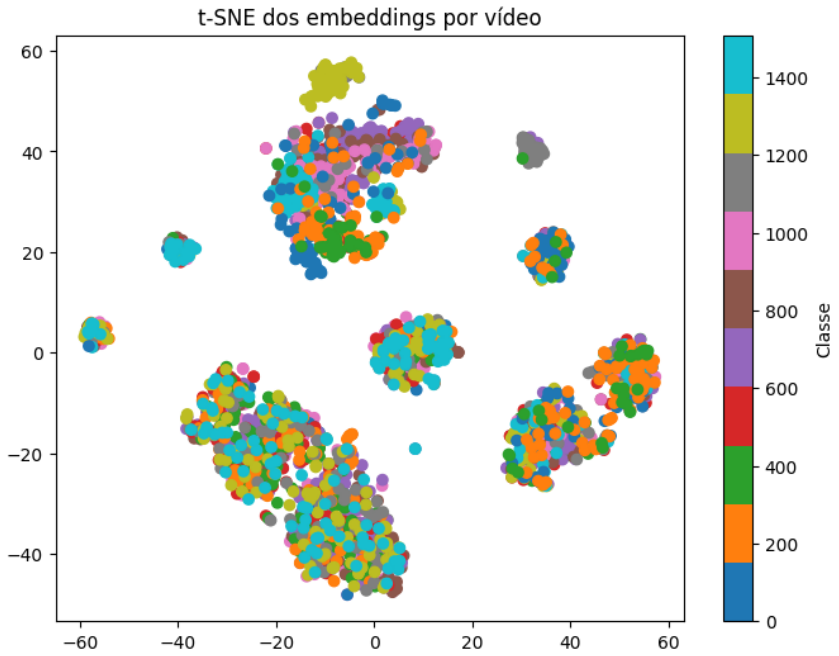


Figure 2. Visualização dos embeddings por vídeo com t-SNE. Cada ponto representa um vídeo, colorido conforme sua classe.

2.1.3. Aplicação utilizando os Embeddings Rede Neural

Treinamos um modelo de regressão logística simples com os vetores extraídos como entrada. A acurácia obtida no conjunto de teste foi de apenas 0.0039, o que pode confirmar que os vetores gerados ainda não carregam informações discriminativas suficientes. Tornando difícil a classificação supervisionada eficaz sem deep learning.

2.1.4. Aplicação utilizando os Embeddings Combinação de Descritores

O objetivo da aplicação dos embeddings é extrair descritores representativos das imagens dos vídeos em LIBRAS utilizando a rede neural pré-treinada ResNet50. O processo envolve o carregamento das imagens de três bases de vídeos, cada uma organizada em diretórios, e a aplicação do modelo ResNet50 para gerar os descritores de cada frame.

A extração dos descritores é realizada através de uma função personalizada que percorre os diretórios e carrega as imagens. Cada imagem é pré-processada, redimensionada para 224x224 pixels e normalizada utilizando o método preprocess$_{i}nputdaResNet50, garantindoqueosdadosestejamadequadosaomodelo.Apósisso, afunçãoe$

O conjunto de descritores de todas as imagens é combinado em um único array, onde cada linha representa o descritor de uma imagem. Esse array é então armazenado em um arquivo CSV, permitindo que os descritores sejam utilizados em etapas subsequentes, como redução de dimensionalidade ou treinamento de modelos de aprendizado de máquina.

2.1.5. Visualização Combinação de Descritores

Após a extração e combinação dos descritores, foi aplicada a técnica de redução de dimensionalidade, utilizando t-SNE e MDS, para representar as amostras em um espaço bidimensional. Essas abordagens possibilitaram a visualização da distribuição das amostras, facilitando a interpretação dos dados e a identificação de padrões ou agrupamentos nas representações das imagens. A visualização obtida torna mais clara a estrutura dos descritores em um formato acessível e intuitivo.

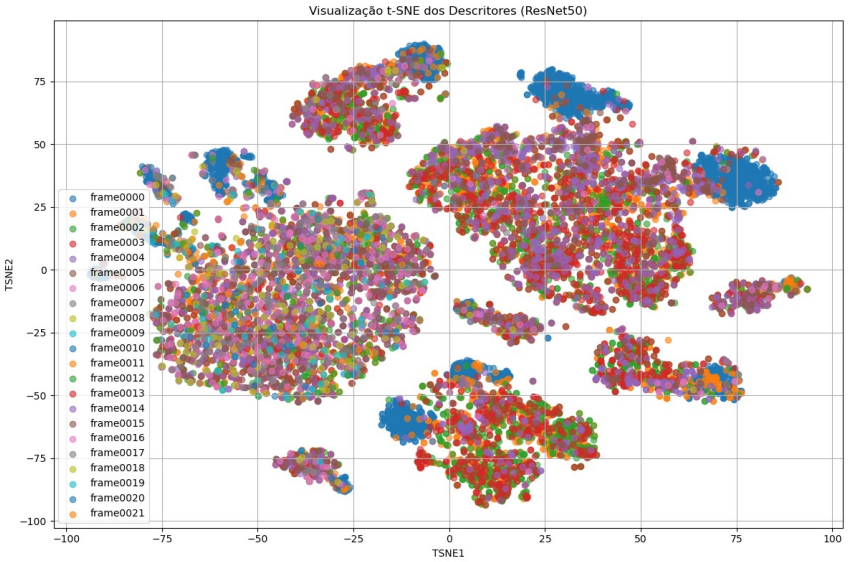


Figure 3. Visualização dos embeddings por descritor com t-SNE.

2.1.6. Análise da Qualidade dos Embeddings Combinação de Descritores

Após a extração dos descritores das imagens utilizando a arquitetura ResNet50, foi aplicada a técnica de redução de dimensionalidade t-SNE, juntamente com o MDS, para representar visualmente os dados em um espaço bidimensional. A visualização revelou a formação de aglomerados bem definidos, sugerindo que os descritores gerados são capazes de capturar características visuais consistentes das imagens, mesmo com variações nas pessoas que realizam os sinais.

No entanto, apesar de algumas separações visíveis, também foi observada uma sobreposição significativa entre os grupos. Isso é esperado, pois os nomes dos arquivos não indicam diretamente a classe dos sinais, mas sim diferentes instâncias do mesmo sinal realizadas por pessoas distintas. Essa sobreposição indica que a separação entre as classes ainda pode ser desafiadora, especialmente quando se considera a variação dos articuladores.

Embora a separação das classes não seja totalmente clara, a visualização sugere que o modelo possui um bom potencial para discriminar entre diferentes sinais quando as classes forem corretamente mapeadas. A análise também evidenciou a robustez do modelo, que parece ser capaz de generalizar bem, mantendo consistência nas representações das imagens, independentemente das variações entre os articuladores.

3. First Page

4. CD-ROMs and Printed Proceedings

5. Sections and Paragraphs

5.1. Subsections

6. Figures and Captions

7. References

References

[Chaveiro et al. 2014] Chaveiro, N., Duarte, S. B. R., Freitas, A. R. d., Barbosa, M. A., Porto, C. C., and Fleck, M. P. d. A. (2014). Qualidade de vida dos surdos que se comunicam pela língua de sinais: revisão integrativa. *Interface*, 18(48):101–114.

[Konečný and Hagara 2013] Konečný, J. and Hagara, M. (2013). One-shot-learning gesture recognition using hog-hof features. *Journal of Machine Learning Research*, 15.

[Mota et al. 2014] Mota, V., Perez, E., Maciel, L., Vieira, M., and Gosselin, P. (2014). A tensor motion descriptor based on histograms of gradients and optical flow. *Pattern Recognition Letters*, 39:85–91. Advances in Pattern Recognition and Computer Vision.

[Rastgoo et al. 2021] Rastgoo, R., Kiani, K., and Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794.

[Sandler et al. 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *arXiv [cs.CV]*.

[Zhan 2019] Zhan, F. (2019). Hand gesture recognition with convolution neural networks. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 295–298.

[Zimmerman et al. 1986] Zimmerman, T. G., Lanier, J., Blanchard, C., Bryson, S., and Harvill, Y. (1986). A hand gesture interface device. *SIGCHI Bull.*, 18(4):189–192.