Uso de Técnicas de Deep Learning para o Reconhecimento da Língua 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. *In this work, a model based on LSTM neural networks was developed and evaluated for the recognition of Brazilian Sign Language (LIBRAS) signals, using keypoint vectors extracted via MediaPipe Holistic. Initially, two distinct architectures were compared: an original, more complex model, and an optimized model that introduced additional regularization mechanisms, including global attention layers and data augmentation techniques. The experiments employed data from multiple datasets, aiming to assess the model's generalization capability in inter-dataset scenarios. Although the optimized model exhibited lower accuracy on the training set, its results demonstrated reduced overfitting and greater robustness during validation and testing, achieving a better balance between performance and generalization. Quantitative and qualitative analyses, including t-SNE visualizations and supervised evaluations, highlighted the relevance of the proposed strategies in handling the high gestural variability characteristic of LIBRAS.*

Resumo. *Neste trabalho, foi desenvolvido e avaliado um modelo baseado em redes neurais LSTM para o reconhecimento de sinais da Língua Brasileira de Sinais (LIBRAS), utilizando vetores de keypoints extraídos via MediaPipe Holistic. Inicialmente, duas arquiteturas distintas foram comparadas: um modelo original, mais complexo, e um modelo otimizado, que introduziu mecanismos adicionais de regularização, incluindo camadas de atenção global e técnicas de data augmentation. Os experimentos empregaram dados de múltiplas bases, buscando avaliar a capacidade de generalização em contextos inter-base. Embora o modelo otimizado tenha apresentado menor acurácia no conjunto de treino, seus resultados evidenciaram menor overfitting e maior robustez na validação e no teste, alcançando melhor equilíbrio entre desempenho e generalização. As análises quantitativas e qualitativas, incluindo visualizações via t-SNE e avaliações supervisionadas, demonstraram a relevância das estratégias propostas para lidar com a alta variabilidade gestual característica da LIBRAS.*

1. Contextualização

A comunicação é um dos pilares essenciais da vida humana. Nossa espécie se destaca pela sociabilidade, uma característica intrínseca que precede, em muitos aspectos, o pleno desenvolvimento da racionalidade complexa. Dentro desse contexto social, a fala emerge como nosso principal método de interação. Contudo, a comunicação vai muito além da

fala; ela se manifesta de diversas maneiras, e para uma parcela considerável da população, essa expressão se dá por meio da língua de sinais.

A língua de sinais é uma esfera universal em que cada país possui sua própria variação. A Organização Mundial de Surdos , dentre os 70 milhões de surdos, coletivamente existam mais de 300 línguas de sinais distintas. 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 à 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. Datasets Utilizados

Neste trabalho foram utilizados três conjuntos de dados principais, selecionados a partir do estudo de referência *A Cross-Dataset Study on the Brazilian Sign Language Translation* (de Avellar Sarmento and Ponti 2023). Os datasets empregados são V-LIBRASIL, Signbank e CEAD, escolhidos devido à sua abrangência e relevância para o reconhecimento automático de sinais da Língua Brasileira de Sinais (LIBRAS).

O dataset V-LIBRASIL é composto por vídeos de sinais realizados por diferentes articuladores, oferecendo diversidade em aspectos como iluminação, fundo e estilo de execução, o que garante maior robustez na generalização dos modelos treinados. Por sua vez, o dataset Signbank, mantido pela Universidade Federal de Santa Catarina (UFSC), constitui um dicionário digital contendo vídeos de sinais padronizados e detalhados, fre-

quentemente utilizado como referência em ensino e pesquisa. Já o dataset CEAD (Centro de Ensino e Apoio à Distância da Universidade Federal de Viçosa - UFV) compreende sinais executados por articuladores específicos, sendo empregado principalmente para avaliação e validação dos modelos desenvolvidos, dada sua estruturação consistente e padronizada.

Ressalta-se que, no presente estudo, apenas o dataset V-LIBRASIL foi utilizado nas etapas de pré-processamento, extração de frames e geração de embeddings visuais. Os datasets Signbank e CEAD foram integrados exclusivamente na fase de aplicação e teste da rede neural treinada, permitindo verificar a capacidade de generalização do modelo sobre bases externas e heterogêneas.

Figure 1. Exemplo do dataset CEAD, ilustrando o sinal “acontecer”.

3. Estudo e Comparação de Embeddings

A etapa de estudo e comparação de embeddings visuais teve como propósito analisar a qualidade das representações geradas para os sinais em LIBRAS, visando entender o quão bem os diferentes sinais são separados no espaço vetorial. Essa análise é essencial para verificar se os embeddings extraídos podem, de fato, ser utilizados como descritores discriminativos para tarefas de reconhecimento ou tradução de sinais.

No presente trabalho, essa etapa foi realizada exclusivamente sobre o dataset V-LIBRASIL. Optou-se por trabalhar somente com esta base nesse estágio porque ela apresenta maior diversidade de articuladores, iluminação e fundos, o que permite avaliar de forma mais robusta a capacidade dos embeddings em capturar variações intra-classe e diferenças inter-classes. Além disso, as outras bases (Signbank e CEAD) foram reservadas para a fase de avaliação cruzada da rede neural treinada, preservando a análise de generalização do modelo.

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.

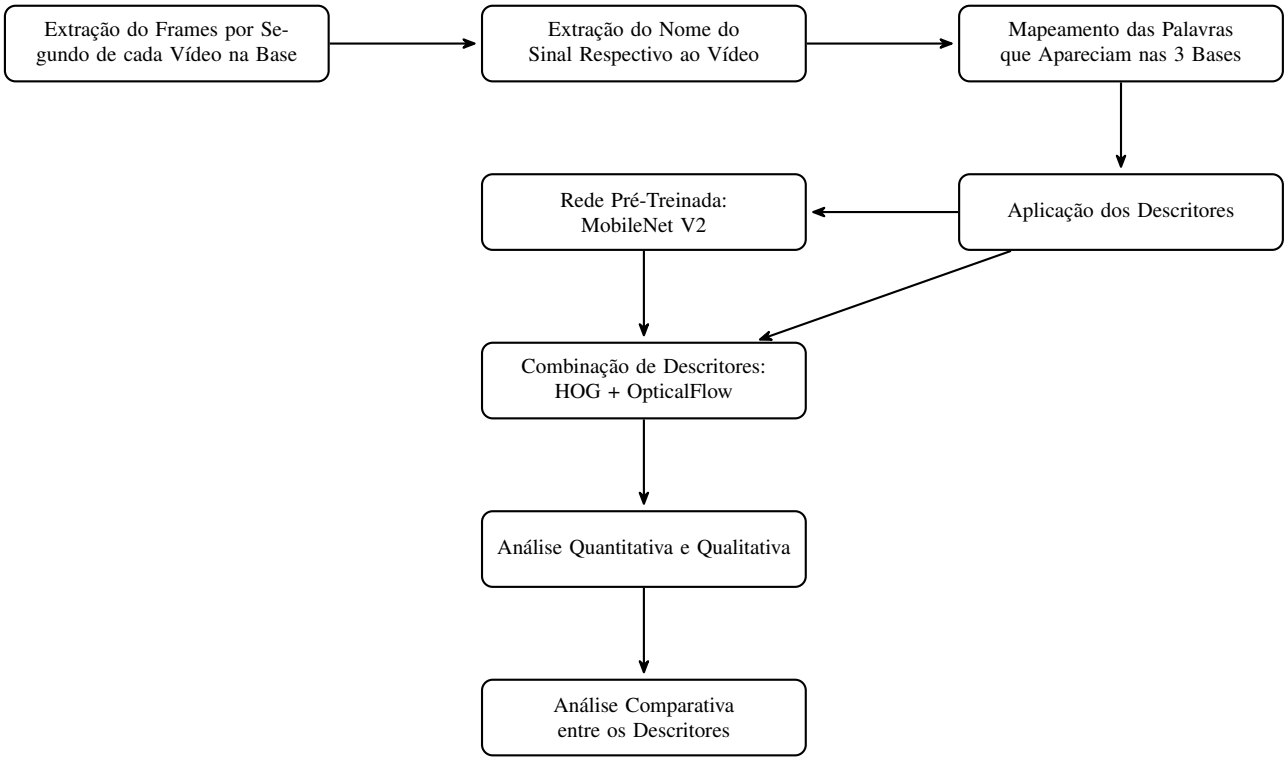

Figure 2. Fluxograma das etapas do processo utilizando os descritores no trabalho.

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

3.1. 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.

Descritor Baseado em uma Rede Neural:

Para a geração de embeddings visuais a partir dos frames extraídos dos vídeos, foi utilizada a arquitetura MobileNetV2, escolhida por seu baixo custo computacional e eficácia comprovada em tarefas de visão computacional. A MobileNetV2, conforme descrita em seu trabalho original (Sandler et al. 2018), possui camadas otimizadas para dispositivos móveis e cenários com restrição de processamento, mantendo excelente desempenho na extração de características visuais.

Neste trabalho, utilizou-se a MobileNetV2 pré-treinada na base ImageNet, aproveitando os pesos já ajustados para a detecção de padrões visuais genéricos. As camadas finais de classificação da rede foram removidas, preservando apenas o extrator convolucional. A saída tridimensional da rede (feature maps) foi então convertida em um vetor unidimensional através da aplicação de uma camada de Global Average Pooling 2D, originando um embedding compacto e representativo para cada frame.

Após a obtenção dos embeddings de todos os frames, foi realizada uma avaliação sistemática da qualidade dessas representações. Três análises complementares foram conduzidas:

- **Cálculo de Similaridade Intra e Interclasse:** foi utilizada a métrica de similaridade cosseno para quantificar a proximidade média entre embeddings de vídeos pertencentes à mesma classe (intra-classe) e entre classes diferentes (inter-classe). No experimento conduzido, observou-se uma similaridade média intra-classe de 0.7936 e inter-classe de 0.8079, evidenciando uma sobreposição considerável entre classes distintas.
- **Visualização com t-SNE:** foi aplicada a técnica t-Distributed Stochastic Neighbor Embedding (t-SNE) para projetar os embeddings em duas dimensões, permitindo análise visual da separação entre diferentes sinais. As projeções revelaram clusters parcialmente sobrepostos, sugerindo que os descritores, embora informativos, não são plenamente discriminativos.
- **Classificação com Regressão Logística:** para avaliar a capacidade discriminativa dos embeddings gerados pela MobileNetV2, foi empregada uma regressão logística como classificador linear. Essa escolha deve-se ao fato de a regressão logística fornecer uma análise simples e interpretável sobre o poder separador dos embeddings, permitindo verificar se os vetores extraídos contêm informações suficientes para distinguir as diferentes classes de sinais. A acurácia obtida per-

maneceu abaixo do esperado, reforçando a percepção de que as representações extraídas não estão perfeitamente adaptadas para capturar as sutilezas semânticas dos sinais em LIBRAS.

Tais análises evidenciam que, embora a MobileNetV2 e as técnicas tradicionais (HOG + Optical Flow) sejam eficazes em capturar aspectos visuais e de movimento, desafios persistem quanto à separação das classes no espaço dos embeddings. Isso destaca a necessidade de abordagens específicas para o domínio de LIBRAS ou estratégias adicionais de refinamento dos descritores extraídos.

3.1.1. Visualização Rede Neural

Com o objetivo de analisar visualmente a distribuição dos embeddings gerados pela rede neural, aplicou-se o algoritmo t-Distributed Stochastic Neighbor Embedding (t-SNE) para reduzir a dimensionalidade dos vetores para duas dimensões. Essa projeção visa identificar padrões de agrupamento ou separação entre as classes no espaço vetorial.

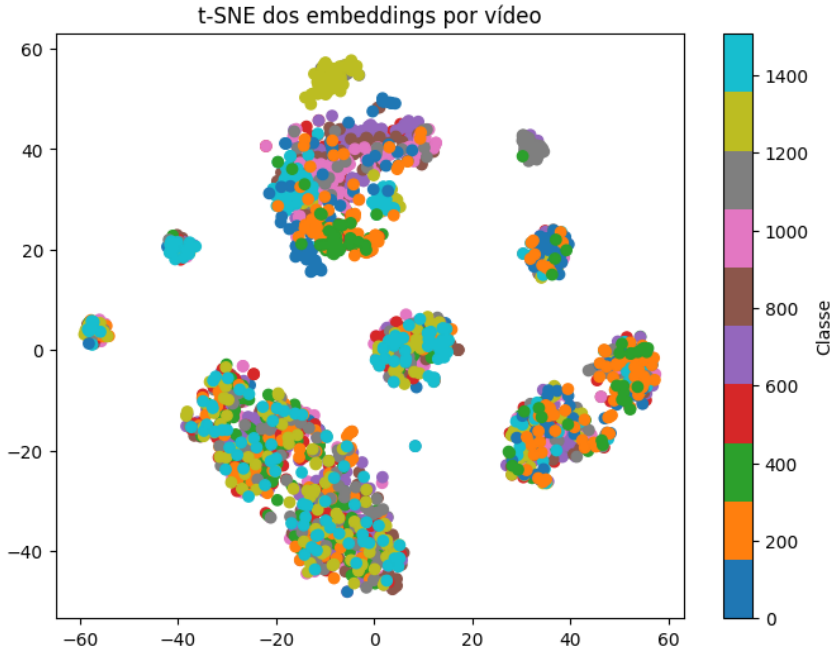

Figure 3. Visualização dos embeddings por vídeo com t-SNE. Cada ponto representa um vídeo, colorido conforme sua classe.

A Figura 3 ilustra o resultado da projeção dos embeddings obtidos a partir da MobileNetV2. Observa-se a formação de alguns agrupamentos visuais aparentes, sugerindo que a rede neural é capaz de capturar certas características discriminativas dos sinais. No entanto, verifica-se também uma considerável sobreposição entre diferentes classes, o que indica limitação na capacidade dos embeddings de separar plenamente as categorias semânticas.

Tal comportamento pode ser justificado pela natureza complexa dos sinais em LIBRAS, que envolvem movimentos sutis, variações entre articuladores e contextos visuais

diversificados. Além disso, como o treinamento da MobileNetV2 foi originalmente realizado sobre a base ImageNet, suas representações podem não estar totalmente adaptadas às particularidades dos sinais manuais da língua de sinais brasileira. Assim, embora o descritor capture informações visuais relevantes, ele ainda apresenta restrições em distinguir com alta precisão as diferentes classes, evidenciado tanto pela sobreposição visual no t-SNE quanto pelos resultados quantitativos obtidos nas análises subsequentes.

3.1.2. Aplicação utilizando os Embeddings Rede Neural

Para avaliar a capacidade discriminativa dos embeddings gerados pela rede MobileNetV2, foi treinado um modelo de regressão logística utilizando os vetores extraídos como variáveis de entrada. O objetivo desse experimento foi verificar se as representações aprendidas pela rede neural seriam suficientemente informativas para permitir a classificação supervisionada das diferentes classes de sinais em LIBRAS, mesmo através de um classificador linear relativamente simples.

No entanto, a acurácia obtida no conjunto de teste foi extremamente baixa, alcançando apenas 0.0039. Esse desempenho indica que os embeddings, apesar de conterem informações visuais relevantes, ainda não apresentam características suficientemente discriminativas para separar adequadamente as diferentes classes do conjunto de dados. Tal resultado evidencia a complexidade inerente aos sinais em LIBRAS, os quais possuem variações sutis entre classes e articuladores, e reforça a limitação de métodos lineares em capturar tais nuances sem o uso de abordagens de aprendizado profundo mais robustas ou estratégias específicas de ajuste dos embeddings ao domínio da língua de sinais brasileira.

3.1.3. Aplicação utilizando os Embeddings da Combinação de Descritores

A proposta de combinação de descritores busca capturar informações complementares presentes nos vídeos, integrando características espaciais e temporais em uma única representação. Para essa finalidade, foram empregados dois descritores tradicionais amplamente utilizados em visão computacional: o Histogram of Oriented Gradients (HOG) e o Optical Flow.

O descritor HOG é responsável por representar a estrutura espacial das imagens, destacando contornos e bordas que caracterizam a forma dos objetos presentes nos frames. Por sua vez, o Optical Flow, calculado pelo método de Farneback, fornece informações sobre o movimento de pixels entre frames consecutivos, permitindo a descrição da dinâmica e da fluidez dos sinais executados nos vídeos, aspecto crucial no contexto da Língua Brasileira de Sinais (LIBRAS).

Cada descritor foi extraído individualmente a partir dos frames dos vídeos. Em seguida, os vetores gerados foram normalizados, garantindo que ambos tivessem escalas comparáveis e evitando que a magnitude de um descritor prevalecesse sobre o outro. Após a normalização, os vetores foram concatenados horizontalmente, originando um único vetor representativo para cada vídeo, capaz de integrar tanto informações estáticas (estruturais) quanto dinâmicas (de movimento).

Devido ao elevado número de variáveis geradas por essa concatenação, foi necessária a aplicação de uma etapa de redução de dimensionalidade utilizando a Análise de Componentes Principais (PCA). Essa técnica teve como objetivo condensar as informações mais relevantes, eliminar redundâncias e tornar os vetores resultantes mais compactos, facilitando as etapas subsequentes de análise e visualização dos dados. Dessa forma, foi possível obter embeddings combinados, capazes de sintetizar de maneira eficiente as principais características espaciais e temporais presentes nos vídeos em LIBRAS.

3.1.4. Visualização Combinação de Descritores

Com os embeddings combinados gerados a partir dos descritores HOG e Optical Flow, foi realizada uma visualização utilizando a técnica de t-Distributed Stochastic Neighbor Embedding (t-SNE). Essa técnica é amplamente empregada para redução de dimensionalidade e projeção de dados de alta dimensão em um espaço bidimensional, facilitando a interpretação visual dos agrupamentos eventualmente formados.

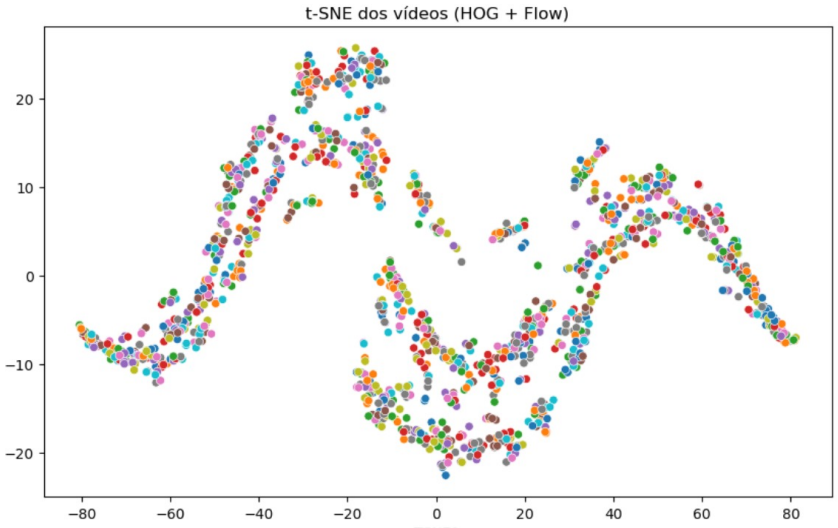

Figure 4. Visualização dos embeddings por HOG x Opticaw Flow com t-SNE.

A projeção dos dados revelou a presença de alguns agrupamentos visuais, que sugerem padrões de similaridade entre os vídeos. No entanto, observou-se uma considerável sobreposição entre diferentes classes, o que dificulta a distinção clara entre elas no espaço projetado. Tal comportamento indica que, mesmo após a fusão dos descritores HOG e Optical Flow e a aplicação do PCA, os vetores gerados não conseguem capturar completamente as diferenças intrínsecas entre as classes representadas nos vídeos.

Esse resultado está alinhado com a hipótese de que, embora a fusão de descritores tradicionais seja rica em informações sobre aspectos espaciais e temporais, ela pode não ser suficientemente discriminativa para garantir, por si só, uma separação clara entre categorias complexas, como aquelas presentes na Língua Brasileira de Sinais (LIBRAS). Dessa forma, a sobreposição visual observada sugere limitações na capacidade dos descritores combinados em representar com precisão as nuances semânticas e articulatórias

dos sinais, apontando para a necessidade de explorar abordagens mais sofisticadas, como redes neurais profundas, para uma melhor discriminação entre classes.

3.1.5. Análise da Qualidade dos Embeddings Combinação de Descritores

A qualidade dos embeddings obtidos pela combinação dos descritores HOG e Optical Flow foi avaliada por meio de métricas quantitativas e análises qualitativas. Inicialmente, foi utilizada a métrica de similaridade cosseno para medir a proximidade entre os vetores pertencentes a vídeos de uma mesma classe (intra-classe) e de classes distintas (inter-classe). Os resultados demonstraram uma média de similaridade intra-classe de -0.0177 e inter-classe de -0.0005, valores próximos de zero ou até ligeiramente negativos, o que indica baixa separabilidade entre os grupos. Tais resultados sugerem que os vetores não estão organizados de maneira suficientemente discriminativa em relação às classes, comprometendo seu potencial como representações informativas para tarefas de reconhecimento.

Complementarmente, foi realizada uma avaliação supervisionada por meio de um classificador de regressão logística treinado sobre os embeddings combinados, com o objetivo de verificar sua capacidade discriminativa em contexto supervisionado. A acurácia obtida no conjunto de teste foi de apenas 0.2105, desempenho consideravelmente abaixo do esperado para um modelo de classificação simples, o que reforça a dificuldade de separar corretamente as amostras das diferentes classes utilizando esses descritores.

Esse resultado evidencia que, embora a fusão dos descritores HOG e Optical Flow consiga integrar diferentes perspectivas da informação visual — contemplando aspectos espaciais e temporais —, os embeddings resultantes não são suficientemente discriminativos para permitir uma classificação precisa entre classes distintas. Tais limitações reforçam a necessidade de explorar abordagens mais robustas, como modelos baseados em aprendizado profundo, capazes de capturar de forma mais eficaz as nuances e variações presentes nos sinais da Língua Brasileira de Sinais (LIBRAS).

4. Treinamento do Modelo

Esta etapa tem como objetivo o treinamento de uma rede LSTM para a tarefa de reconhecimento de sinais da Língua Brasileira de Sinais (LIBRAS), utilizando como entrada os vetores de keypoints extraídos com o MediaPipe Holistic. A configuração atual do conjunto de dados foi definida a partir das análises conduzidas na etapa anterior, dedicada à extração e exploração de embeddings. Naquela fase, observou-se uma baixa separabilidade entre classes e uma elevada variabilidade intra-classe, o que motivou uma subseleção criteriosa das categorias mais consistentes e a coleta de novas instâncias com melhor qualidade visual.

Durante o treinamento, diferentes configurações de hiperparâmetros foram avaliadas de forma sequencial, com base no desempenho obtido em validação. O processo foi conduzido de maneira iterativa, com ajustes sucessivos visando à melhoria da acurácia do modelo, permitindo refinar gradualmente o comportamento da rede ao longo dos ciclos de teste.

4.1. Seleção de Classes

A definição das classes utilizadas no presente experimento foi fundamentada no estudo "A Cross-Dataset Study on the Brazilian Sign Language Translation" (de Avellar Sarmento and Ponti 2023), apresentado no workshop da ICCVW. Esse trabalho teve como principal contribuição a padronização de um benchmark para tarefas de classificação de sinais em LIBRAS, com ênfase na avaliação da capacidade de generalização dos modelos em cenários realistas.

No referido estudo, os autores integraram quatro bases de dados distintas: V-LIBRASIL (UFPE), LIBRAS-Português (UFV) (), Dicionário de LIBRAS (INES) e SignBank (UFSC). Cada base possui características próprias quanto ao número de vídeos, variedade de articuladores e qualidade de gravação, o que impôs o desafio de selecionar um subconjunto de sinais que fosse comum a todas as fontes.

Inicialmente, foram identificadas 49 categorias que estavam presentes em todos os conjuntos e que possuíam, no mínimo, quatro instâncias por base. No entanto, observou-se que, mesmo entre sinais com o mesmo rótulo, havia significativa variação na execução — reflexo de diferenças regionais, estilo individual e qualidade dos vídeos. Para mitigar esse ruído semântico e visual, foi realizada uma análise de variabilidade intra-classe, a partir da qual as categorias com maior inconsistência foram descartadas.

Após esse refinamento, chegou-se a um conjunto mais robusto de 37 classes, selecionadas com base na estabilidade visual dos sinais e na representatividade das instâncias. Essa curadoria garantiu maior homogeneidade nas amostras e favoreceu a avaliação mais justa do modelo.

No presente trabalho, foram utilizadas exclusivamente instâncias do Articulador 2 da base V-LIBRASIL (UFPE), totalizando três vídeos por classe. Complementarmente, foi incluída uma instância por classe proveniente das bases auxiliares (Universidade Federal de Viçosa 2025) CEAD e (Universidade Federal de Santa Catarina 2025) SignBank, totalizando 185 amostras. Essa escolha teve como objetivo explorar variações interbase de forma controlada, mantendo a coerência com o benchmark proposto no estudo de referência.

Table 1. Tabela de instâncias utilizadas no treinamento

Base	Origem	Instâncias por classe	Total de instâncias
V-LIBRASIL (Articulador 2)	UFPE	3	111
CEAD	UFV	1	44
Signbank	UFSC	1	37
Total	—	1 a 3	37

Entre as classes selecionadas, encontram-se sinais que representam objetos concretos, ações do cotidiano e conceitos diversos, tais como: acordar, amigo, bicicleta, coelho, comer, comparar, escola, esquecer, flor, melancia e patins, entre outros. A escolha dessas categorias visa avaliar a capacidade do modelo em discriminar gestos visualmente semelhantes, bem como sua sensibilidade a variações na pose corporal, expressões faciais e configurações manuais — aspectos fundamentais na estruturação dos sinais em LIBRAS.

4.2. Extração de Keypoints

A extração de características dos vídeos foi realizada utilizando o modelo Holistic da biblioteca MediaPipe, que permite a detecção simultânea de pontos de referência na face (468), mãos (21 para cada) e corpo (33). Cada frame é convertido em um vetor de 1662 dimensões, resultante da concatenação dos pontos tridimensionais (x, y, z) e, no caso da pose, também a visibilidade. As imagens são processadas a uma taxa de 1 frame por segundo, e os vetores correspondentes são armazenados sequencialmente em arquivos .npy para posterior uso como entrada na rede.

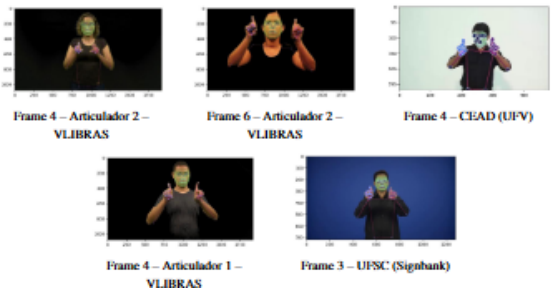

Figure 5. Amostras extraídas das bases utilizadas no experimento, anotadas com pontos-chave (*keypoints*) extraídos pelo modelo MediaPipe Holistic.

4.3. Definição do Modelo LSTM

Após o refinamento prévio dos dados, no qual se definiu um conjunto robusto e representativo de 37 classes para reduzir ambiguidades e garantir maior consistência visual, avançou-se para a definição e implementação da arquitetura do modelo de reconhecimento de sinais da Língua Brasileira de Sinais (LIBRAS).

A escolha recaiu sobre uma rede do tipo Long Short-Term Memory (LSTM), uma variante das redes neurais recorrentes (RNNs) projetada para processar dados sequenciais. Tal escolha se justifica pelas características temporais dos sinais em LIBRAS, que envolvem sequências coordenadas de movimentos das mãos, expressões faciais e postura corporal ao longo do tempo.

A arquitetura definida foi concebida para receber como entrada sequências temporais de vetores de keypoints extraídos dos vídeos, contemplando informações tridimensionais das poses corporais e faciais. Além disso, foram incorporadas técnicas de normalização dos dados, camadas densas para extração de padrões não lineares, e mecanismos de regularização para mitigar overfitting, garantindo maior capacidade de generalização do modelo.

As etapas subsequentes detalham a configuração específica do modelo implementado, bem como as estratégias de treinamento e avaliação adotadas neste trabalho.

4.4. Configuração do Modelo

O modelo desenvolvido neste trabalho é baseado em uma arquitetura do tipo LSTM (Long Short-Term Memory), uma variante de redes neurais recorrentes (RNNs) especialmente projetada para lidar com dados sequenciais e temporais. A escolha dessa arquitetura foi inspirada no estudo de (de Avellar Sarmento and Ponti 2023), que investigou abordagens

de reconhecimento de sinais em LIBRAS e destacou a importância de modelos capazes de capturar padrões temporais em dados gestuais.

As redes LSTM utilizam células especiais capazes de controlar o fluxo de informações ao longo do tempo, permitindo que o modelo "lembre" ou "esqueça" informações relevantes em sequências longas. Tal mecanismo é viabilizado pelo uso das chamadas "portas" (gates), que regulam a atualização e retenção dos estados internos da rede, mitigando problemas como o desaparecimento ou explosão do gradiente, frequentemente encontrados em RNNs tradicionais.

Assim, a arquitetura LSTM mostrou-se particularmente adequada ao reconhecimento de LIBRAS, uma vez que os sinais são compostos por sequências temporais de movimentos das mãos, rosto e corpo, exigindo do modelo a habilidade de capturar não apenas padrões espaciais, mas também a dinâmica temporal subjacente aos gestos. O esquema geral da arquitetura é apresentado na Figura **??**.

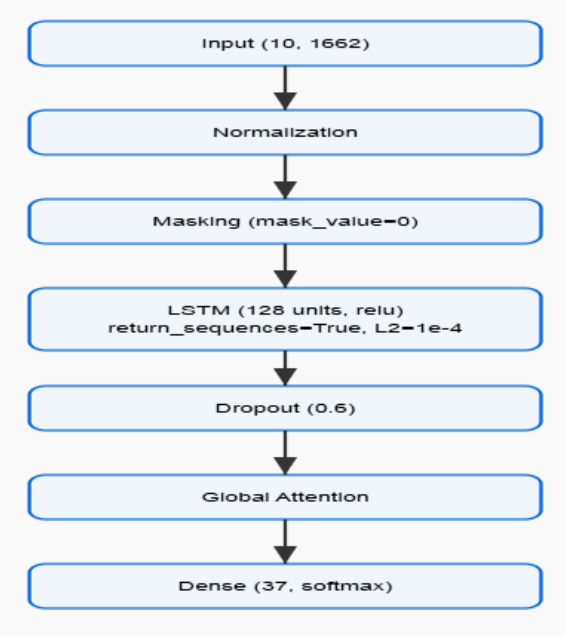

Figure 6. Diagrama da arquitetura do modelo LSTM desenvolvido neste trabalho.

Entrada e Pré-processamento: O modelo recebe como entrada sequências temporais compostas por 10 frames, sendo cada frame representado por um vetor de 1662 atributos correspondentes às coordenadas tridimensionais dos keypoints corporais, faciais e das mãos, extraídos via MediaPipe Holistic. Antes de serem processados, os dados passam por uma camada de normalização e por uma camada de mascaramento, a qual ignora frames nulos na sequência.

Camada LSTM: A etapa subsequente é composta por uma camada LSTM com 128 unidades e ativação ReLU, configurada para retornar sequências. A regularização L2 com coeficiente 10^{-4} é aplicada para mitigar overfitting, dado o alto número de parâmetros e a variabilidade intrínseca dos dados gestuais.

Dropout: Para reduzir ainda mais o risco de sobreajuste, aplica-se uma camada

de Dropout com taxa de 60%.

Mecanismo de Atenção Global: Sobre as sequências geradas pela LSTM, implementou-se uma camada de atenção global personalizada. Esta camada aprende pesos que indicam a importância relativa de cada timestep na sequência, realizando uma soma ponderada para produzir um vetor único representativo do vídeo completo. Essa abordagem visa concentrar o aprendizado nos frames mais informativos, aumentando a robustez do modelo.

Camada de Saída: Finalmente, aplica-se uma camada densa totalmente conectada com 37 unidades e ativação softmax, compatível com o número de classes definidas após o processo de curadoria das categorias de sinais.

Otimização: O modelo é treinado utilizando o otimizador AdamW, com taxa de aprendizado de 10^{-4} e decaimento de pesos igual a 10^{-3}, utilizando a função de perda categorical crossentropy e métricas de avaliação de acurácia e Top-5 Accuracy.

Além disso, durante o treinamento, foram implementadas técnicas de *data augmentation* sobre os vetores de keypoints, incluindo adição de ruído gaussiano, dropout aleatório de keypoints e pequenas rotações, a fim de ampliar a diversidade do conjunto de dados e fortalecer a capacidade do modelo de generalizar para variações na execução dos sinais.

4.5. Processo de Treinamento e Ajustes de Arquitetura

Foram conduzidos experimentos comparativos entre duas arquiteturas LSTM para o reconhecimento de sinais em LIBRAS, a fim de avaliar o impacto das configurações no aprendizado e na generalização do modelo.

O primeiro modelo utilizou 512 unidades LSTM, regularização L1 ($\lambda = 0{,}001$), dropout de 40% e camada densa com 38 saídas. Apesar do bom desempenho no treino, apresentou forte sobreajuste (*overfitting*) e dificuldades na generalização para bases externas.

Para mitigar esse problema, foi proposta uma nova arquitetura, reduzindo a LSTM para 128 unidades, adotando regularização L2 (1×10^{-4}), aumento do dropout para 60% e inclusão de uma camada de atenção global, visando maior foco em padrões temporais relevantes. Também foram aplicadas técnicas de *data augmentation* diretamente sobre os keypoints, como ruído gaussiano, dropout e pequenas rotações.

Embora a nova configuração tenha apresentado leve redução na acurácia geral, demonstrou maior estabilidade e menor diferença entre treino e validação, indicando menor tendência ao sobreajuste e melhor potencial de generalização.

Os resultados comparativos de desempenho entre as duas abordagens são discutidos na subseção seguinte.

4.6. Resultados e Análise Crítica

Nesta subseção, apresentam-se os resultados obtidos no treinamento e no teste dos modelos antigo e novo desenvolvidos para o reconhecimento de sinais em LIBRAS. O objetivo é comparar o desempenho das arquiteturas quanto à capacidade de generalização, estabilidade de treino e eficácia na tarefa de classificação multiclasse.

Desempenho no Conjunto de Treino

Durante o treinamento, foram monitoradas as métricas de acurácia e perda (loss) em cada época, possibilitando acompanhar a evolução do aprendizado dos modelos.

O modelo antigo apresentou sua melhor performance no conjunto de treino na época 56, alcançando uma acurácia de 40,69%, com perda (loss) de 3,0789 e acurácia top-5 de 82,22%. Apesar dos valores relativamente altos no treino, observou-se um comportamento típico de overfitting, evidenciado pela elevação do loss de validação após as primeiras épocas e pela divergência entre as curvas de treino e validação.

Por sua vez, o modelo novo obteve seu melhor resultado no treino na época 70, registrando uma acurácia de 22,22% e perda (loss) de 5,5707. Embora sua acurácia de treino tenha sido inferior à do modelo antigo, o novo modelo demonstrou curvas de loss de validação consideravelmente mais estáveis e com valores mais baixos ao longo das épocas, indicando menor sobreajuste e melhor capacidade de generalização.

De forma geral, a acurácia de treino do modelo novo cresceu de maneira mais gradual e controlada, enquanto o modelo antigo rapidamente atingiu altos valores, mas não sustentou o desempenho na validação. Já o loss de validação do modelo novo manteve-se em patamares mais baixos e estáveis, reforçando sua robustez frente a dados não vistos.

A Figura 7 ilustra o comportamento comparado dos modelos em termos de acurácia de treino e perda de validação, evidenciando as diferenças entre as duas abordagens.

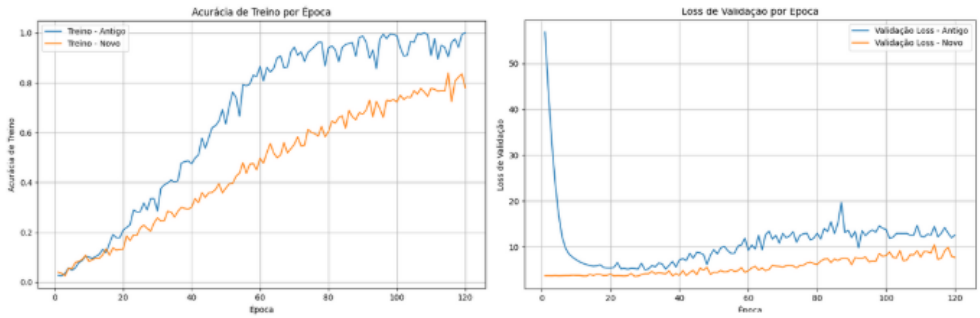

Figure 7. Comparação entre as curvas de acurácia de treino (à esquerda) e loss de validação (à direita) para os modelos antigo e novo.

Desempenho no Conjunto de Teste

A Tabela 2 apresenta os melhores resultados obtidos por cada modelo no conjunto de teste. O **modelo novo** superou o modelo antigo em termos de perda (loss), reduzindo o valor de 12,7589 para 7,6795. Houve também um aumento na acurácia do teste, que passou de 13,51% no modelo antigo para 20,83% no modelo novo. Ainda que a top-5 accuracy do novo modelo (33,33%) tenha sido levemente inferior à do antigo (37,84%), a redução significativa na perda e a melhora da acurácia geral apontam para um modelo mais robusto e menos sujeito a overfitting.

Table 2. Comparação dos melhores resultados obtidos pelos modelos antigo e novo no conjunto de teste

Modelo	Loss (Teste)	Acc (Teste)	Top-5 Acc (Teste)
Antigo	12,7589	13,51%	37,84%
Novo	7,6795	20,83%	33,33%

Análise Crítica dos Resultados

A análise dos resultados evidencia que o **novo modelo** apresenta melhor capacidade de generalização em relação ao modelo antigo, sobretudo pelos valores mais baixos de perda no conjunto de teste. A redução no loss sugere que o modelo está ajustando-se de forma mais adequada às características dos dados externos, mesmo diante do desafio de trabalhar com dados de bases heterogêneas.

Apesar de ter uma acurácia top-5 ligeiramente inferior, o novo modelo demonstra menor overfitting, como visto na evolução mais estável das métricas durante o treinamento. Essa estabilidade e a diminuição da diferença entre as métricas de treino e teste apontam para um modelo mais robusto para aplicações práticas. O aumento de complexidade no novo modelo, com o uso de regularização L1, maior dropout e atenção global, contribuiu para mitigar o sobreajuste, ainda que tenha exigido mais épocas para convergir.

De forma geral, conclui-se que o novo modelo se mostrou superior, principalmente por reduzir a perda em teste e melhorar a acurácia geral, representando um avanço significativo na tarefa de reconhecimento automático de sinais em LIBRAS.

Conclusão

Este trabalho teve como objetivo investigar e aprimorar técnicas para o reconhecimento automático de sinais da Língua Brasileira de Sinais (LIBRAS), utilizando redes neurais profundas aplicadas a dados de keypoints corporais extraídos de vídeos. Inicialmente, revisou-se a literatura sobre métodos de extração e análise de embeddings, destacando-se as dificuldades inerentes à natureza gestual da LIBRAS, como alta variabilidade entre articuladores, diferenças nas condições de gravação e grande semelhança entre determinados sinais.

Foram conduzidos experimentos comparativos entre duas arquiteturas distintas de redes LSTM: uma configuração original, mais complexa, e uma versão otimizada, que incorporou mecanismos adicionais de regularização, atenção global e técnicas de *data augmentation*. A análise das curvas de treino e validação, bem como das métricas quantitativas nos conjuntos de teste, evidenciou que, apesar de o modelo original ter apresentado elevada acurácia no treino, sofreu fortemente com *overfitting*, resultando em menor capacidade de generalização.

Em contrapartida, o modelo otimizado demonstrou menor perda (*loss*) e maior estabilidade durante o treinamento, alcançando melhores resultados no conjunto de teste. Ainda que sua acurácia de treino tenha sido inferior, a redução do *overfitting* e o desempenho mais robusto na validação indicaram melhor equilíbrio entre aprendizado e generalização. As análises com técnicas de visualização, como t-SNE, confirmaram a dificuldade na separação perfeita das classes, mas também mostraram ganhos importantes

na organização espacial dos embeddings no modelo revisado.

Além dos aspectos técnicos, o trabalho reforça a complexidade da tarefa de reconhecimento de LIBRAS, especialmente em cenários interbase, onde as diferenças entre articuladores, cenários e estilos de execução dos sinais impõem desafios adicionais. Ficou claro que, para aplicações práticas, não basta apenas treinar modelos com altas taxas de acerto no treino; é fundamental desenvolver sistemas capazes de manter desempenho consistente frente a dados nunca vistos.

Como perspectiva futura, sugere-se explorar arquiteturas híbridas que combinem modelos recorrentes e convolucionais, integrar mecanismos mais avançados de atenção, como Transformers, e investigar abordagens auto-supervisionadas para enriquecer os embeddings sem necessidade de rotulações extensivas. Além disso, estudos com bases mais amplas e diversificadas poderão contribuir para sistemas de tradução de LIBRAS mais precisos e inclusivos, promovendo avanços significativos na comunicação acessível para a comunidade surda.

Em suma, os resultados obtidos neste estudo representam um passo importante rumo à construção de modelos mais eficientes e robustos para o reconhecimento automático de sinais em LIBRAS, reforçando a importância de técnicas de regularização e simplicidade arquitetural para enfrentar os desafios deste domínio tão complexo e relevante.

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.

[de Avellar Sarmento and Ponti 2023] de Avellar Sarmento, A. H. and Ponti, M. A. (2023). A cross-dataset study on the brazilian sign language translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. Disponível em: `https://github.com/avellar-amanda/LIBRAS-Translation/`. Acesso em: 8 jun. 2025.

[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]*.

[Universidade Federal de Santa Catarina 2025] Universidade Federal de Santa Catarina (2025). Signbank – dicionário de língua brasileira de sinais. Disponível em: `https://signbank.libras.ufsc.br/pt/`. Acesso em: 8 jun. 2025.

[Universidade Federal de Viçosa 2025] Universidade Federal de Viçosa (2025). Dicionário de libras – projeto inovar mais. Disponível em: `https://sistemas.cead.ufv.br/capes/dicionario/`. Acesso em: 8 jun. 2025.

[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.