Explorando a Integração de Múltiplas Modalidades Visuais no Reconhecimento Contínuo de Libras

Guilherme Galvão de Oliveira Pinto[1], **Guilherme Vieira Leite**[1]
Hélio Pedrini[2], **José Mario De Martino**[1]

[1]Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas

[2]Instituto de Computação
Universidade Estadual de Campinas

{g258485, g210404}@dac.unicamp.br, {pedrini, martino}@unicamp.br

Abstract. *Continuous Brazilian Sign Language (Libras) recognition presents challenges related to the multimodal nature of signing and the modeling of complex spatio-temporal dependencies. This ongoing work investigates the adaptation of neural architectures for continuous Libras recognition using RGB videos, body, facial, and hand keypoints, as well as multi-tier linguistic annotations produced in ELAN. The proposed approach combines visual feature extraction, multimodal integration, and temporal modeling. Evaluation will be conducted using the Word Error Rate (WER) metric and ablation studies. The goal is to analyze the impact of multimodality and multi-layer annotations on the performance of continuous Libras recognition models.*

Resumo. *O reconhecimento contínuo de Língua Brasileira de Sinais (Libras) apresenta desafios relacionados à natureza multimodal da sinalização e à modelagem de dependências espaço-temporais complexas. Este trabalho em andamento busca investigar a adaptação de arquiteturas neurais para o reconhecimento contínuo de Libras utilizando vídeos RGB, keypoints corporais, faciais e manuais, além de anotações linguísticas multi-tier produzidas no ELAN. O método proposto combina extração de características visuais, integração multimodal e modelagem temporal. A avaliação será realizada por meio da métrica Word Error Rate (WER) e estudos de ablação. Espera-se analisar o impacto da multimodalidade e das anotações multicamadas no desempenho de modelos de reconhecimento contínuo de Libras.*

1. Introdução

A Língua Brasileira de Sinais (Libras) constitui o principal meio de comunicação da comunidade surda brasileira e desempenha papel fundamental na promoção da acessibilidade e da inclusão social [Presidência da República 2002, Presidência da República 2015]. Diferentemente da língua portuguesa, a Libras utiliza simultaneamente componentes manuais e não manuais, incluindo configurações de mão, movimentos corporais, expressões faciais, direção do olhar e organização espacial da mensagem. Essa natureza visual-espacial e multimodal torna o desenvolvimento de tecnologias para processamento automático de línguas de sinais um desafio científico

relevante e uma importante ferramenta para ampliar a acessibilidade em diferentes contextos sociais, educacionais e institucionais [Bragg et al. 2019].

Apesar da relevância dessas tecnologias, o reconhecimento contínuo da língua de sinais ainda é um desafio [Khan et al. 2025, Najib 2025, Wang et al. 2025]. Diferentemente de sinais isolados, reconhecê-los exige lidar com sequências de sinais de forma contínua, envolvendo fenômenos como coarticulação, transições entre sinais, variações de velocidade e a integração de múltiplos articuladores visuais. Como discutem Koller et al. [Koller et al. 2015], há uma demanda de modelos capazes de capturar dependências espaço-temporais complexas e de operar em condições mais próximas das situações reais de uso. Nos últimos anos, a literatura tem apresentado diversos métodos que buscam solucionar esses problemas. Por exemplo, Camgöz et al. [Camgoz et al. 2018] propuseram uma formulação neural para a tarefa, enquanto Camgöz et al. [Camgoz et al. 2020b] apresentaram uma arquitetura baseada em *Transformers* para reconhecimento e tradução conjuntos. Além disso, Camgöz et al. [Camgoz et al. 2020a], Yin e Read [Yin and Read 2020] e Yin et al. [Yin et al. 2023] exploraram múltiplos articuladores e novas formas de incorporar informação linguística intermediária aos modelos.

Apesar desses avanços, ainda há muitas questões em aberto, especialmente no contexto da Libras. Este trabalho está inserido em um projeto mais amplo voltado ao reconhecimento e à tradução de língua de sinais, envolvendo o desenvolvimento de um corpus multimodal de Libras contendo vídeos RGB, *keypoints* corporais, faciais e manuais, além de anotações linguísticas *multi-tier* produzidas no ELAN. O ELAN é uma ferramenta de anotação linguística que permite registrar e analisar, de forma sincronizada ao vídeo, produções em línguas de sinais [Wittenburg et al. 2006]. Este trabalho tem como objetivo principal avaliar e adaptar arquiteturas neurais para o reconhecimento contínuo, investigando o impacto de diferentes modalidades visuais e estratégias de fusão multimodal, de modo a responder à seguinte questão: como integrar, de forma consistente, vídeos RGB, *keypoints* e anotações linguísticas *multi-tier* para melhorar o desempenho de modelos neurais de reconhecimento contínuo de Libras?

2. Metodologia

A metodologia proposta visa investigar o impacto de diferentes modalidades visuais e anotações linguísticas multicamadas no reconhecimento contínuo de Libras. Três tarefas principais compõem a metodologia: construção de corpus multimodal, extração e integração de representações multimodais e modelagem temporal baseada em arquiteturas neurais.

2.1. Construção de Corpus Multimodal de Libras

Os experimentos são conduzidos com um corpus multimodal de Libras, em desenvolvimento no âmbito de um projeto de pesquisa mais amplo. O corpus é composto por sentenças em português traduzidas para Libras por uma equipe de sete intérpretes, incluindo três participantes surdos. As gravações são realizadas em ambiente controlado utilizando quatro câmeras RGB posicionadas em diferentes ângulos. Essa configuração reduz problemas de oclusão e permite a obtenção de múltiplas perspectivas da sinalização. A partir dos vídeos, são extraídas representações complementares, incluindo *keypoints*

corporais, faciais e das mãos, possibilitando a construção de diferentes modalidades de entrada para os modelos de reconhecimento.

Além das informações visuais, o corpus incorpora anotações linguísticas produzidas no ELAN [Wittenburg et al. 2006]. Diferentemente de abordagens baseadas em uma única camada de anotação, a estrutura adotada utiliza múltiplos *tiers* sincronizados temporalmente. Além das glosas, são registradas informações sobre apontamentos espaciais, mão dominante, direção do olhar e marcadores não manuais. Essa organização permite uma representação mais detalhada da sinalização e fornece informações adicionais potencialmente úteis para o treinamento de modelos neurais. Atualmente, o corpus abrange diferentes domínios de aplicação, incluindo contextos de saúde, alimentação e mobilidade, o que contribui para uma maior diversidade linguística e contextual dos dados.

2.2. Extração e Integração de Representações Multimodais

A primeira etapa do processo computacional consiste na conversão de vídeos em representações adequadas ao treinamento de modelos de reconhecimento. Os vídeos RGB são processados por extratores de características visuais, que geram *embeddings* espaciais para cada quadro da sequência. Paralelamente, representações estruturadas derivadas dos vídeos, como *keypoints* corporais, faciais e manuais, são convertidas em vetores numéricos que descrevem explicitamente a configuração espacial dos articuladores envolvidos na sinalização. Dessa forma, o reconhecimento deixa de operar diretamente sobre o vídeo bruto e passa a utilizar sequências de representações compactas que preservam informações relevantes para a modelagem da sinalização.

Um dos objetivos centrais deste trabalho é investigar como diferentes fontes de informação podem ser combinadas para apoiar o reconhecimento contínuo de Libras. Para isso, serão avaliadas diferentes estratégias de fusão multimodal. Na abordagem de fusão antecipada (*early fusion*), as representações provenientes das diferentes modalidades são combinadas antes da modelagem temporal. Já na fusão tardia (*late fusion*), cada modalidade é inicialmente processada por um fluxo independente, sendo integrada apenas em estágios posteriores da rede. A comparação entre essas estratégias permitirá avaliar em que medida as diferentes modalidades contribuem para o desempenho do sistema.

2.3. Modelagem Temporal para Reconhecimento Contínuo

O reconhecimento contínuo está sendo explorado por meio de arquiteturas neurais baseadas em *Transformers*, tomando como referência os modelos propostos por Camgöz et al. (2020) [Camgoz et al. 2020b]. A escolha dessa arquitetura se justifica por seu papel de destaque na literatura da área, tanto pelos resultados reportados quanto por oferecer uma formulação adequada para modelar sequências visuais longas, com dependências espaço-temporais complexas, características centrais do reconhecimento contínuo. A Figura 1 sintetiza as principais etapas da arquitetura, desde a obtenção das representações espaciais até a predição da sequência de glosas e a supervisão do treinamento.

Após a etapa de fusão, as sequências de representações multimodais são processadas por mecanismos de atenção capazes de modelar dependências de longo alcance ao longo da sinalização. Essa etapa busca capturar fenômenos característicos do reconhecimento contínuo, incluindo coarticulação, variações entre sinalizantes e relações contextuais entre sinais consecutivos. Como saída, o modelo produz uma sequência de glosas correspondente ao conteúdo linguístico presente no vídeo de entrada.

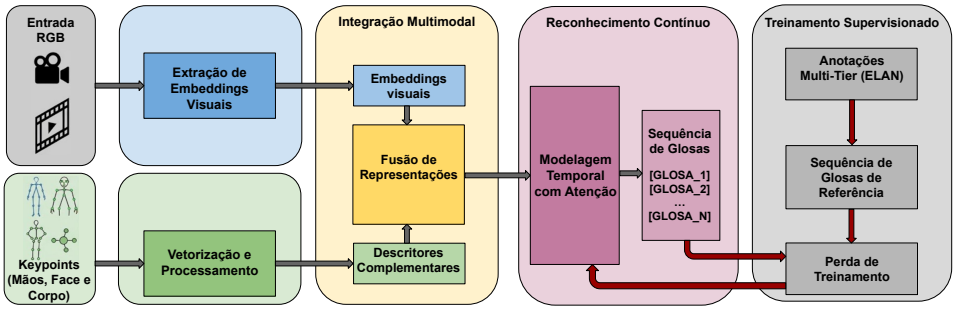

Figura 1. Fluxo computacional para o reconhecimento contínuo de Libras. As setas vermelhas indicam o uso de anotações *multi-tier* na supervisão do treinamento.

3. Avaliação Experimental

A avaliação do sistema será conduzida por meio da comparação entre as sequências de glosas produzidas pelo modelo e as anotações de referência presentes no corpus. Como métrica principal, será utilizada a *Word Error Rate* (WER), amplamente empregada em tarefas de reconhecimento contínuo de língua de sinais [Camgoz et al. 2020b]. A WER quantifica a distância de edição entre a sequência predita e a correta, considerando inserções, remoções e substituições. Além da avaliação global, serão realizados estudos de ablação com o objetivo de mensurar a contribuição individual de cada modalidade visual e das anotações multicamadas para o desempenho final do sistema.

4. Considerações Finais

Um dos principais desafios da pesquisa reside na reprodutibilidade técnica das arquiteturas de referência. A adaptação do modelo descrito em [Camgoz et al. 2020b] exige lidar com dependências de códigos, bibliotecas desatualizadas e escolhas de implementação nem sempre totalmente especificadas. Além disso, há um gargalo importante na extração de características espaciais [Alvarez et al. 2022], pois esse tipo de modelo não opera diretamente sobre o vídeo bruto, mas sobre *features* previamente extraídas. Como a arquitetura exata do extrator original não foi disponibilizada integralmente, torna-se necessário investigar alternativas para gerar *embeddings* visuais robustos, etapa central para o restante da tarefa de reconhecimento. Outro desafio diz respeito à generalização do modelo para o cenário de Libras. O sistema de referência foi desenvolvido sobre o conjunto de dados PHOENIX14T, restrito ao domínio de previsões do tempo e à Língua de Sinais Alemã, enquanto o presente trabalho se concentra na Libras e abrange domínios linguísticos mais variados.

A principal contribuição deste trabalho reside na consolidação de uma base metodológica para a adaptação e avaliação de arquiteturas neurais voltadas a esse problema, com a definição de um fluxo computacional que envolve extração de características espaciais, fusão multimodal, modelagem temporal e avaliação por WER. Embora ainda não haja resultados experimentais finais, o mapeamento dos desafios de reprodutibilidade, a extração de *features* e a generalização para o contexto de Libras já representam etapas relevantes da pesquisa. Espera-se, assim, que os próximos experimentos avancem na compreensão do impacto da multimodalidade e das anotações *multi-tier* no reconhecimento contínuo, contribuindo para o amadurecimento metodológico da área e para o desenvolvimento de tecnologias mais robustas e aplicáveis à comunidade surda brasileira.

Agradecimentos

Este trabalho é financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Processo 2026/05600-6 (Bolsa de Iniciação Científica), Processo 2025/12630-6 (Bolsa de Pós-Doutorado) e Processo 2024/00914-7 (Centro de Ciência para o Desenvolvimento: Tecnologia Assistiva e Acessibilidade em Libras CCD-TAAL).

Referências

Alvarez, P. C., Nieto, X. G., and Benet, L. T. (2022). Sign Language Translation Based on Transformers for the How2Sign Dataset. *Image Processing Group Signal Theory and Communications Department Universitat Politècnica de Catalunya*.

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al. (2019). Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020a). Multi-Channel Transformers for Multi-Articulatory Sign Language Translation. In *European Conference on Computer Vision*, pages 301–319. Springer.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020b). Sign Language Transformers: Joint End-To-End Sign Language Recognition and Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.

Khan, A., Jin, S., Lee, G.-H., Arzu, G. E., Dang, L. M., Nguyen, T. N., Choi, W., and Moon, H. (2025). Deep Learning Approaches for Continuous Sign Language Recognition: A Comprehensive Review. *IEEE Access*, 13:55524–55544.

Koller, O., Forster, J., and Ney, H. (2015). Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers. *Computer Vision and Image Understanding*, 141:108–125.

Najib, F. M. (2025). Sign Language Interpretation using Machine Learning and Artificial Intelligence. *Neural Computing and Applications*, 37(2):841–857.

Presidência da República (2002). Lei número 10.436. Diário Oficial da União, Imprensa Nacional, n. 79, Seção 1.

Presidência da República (2015). Lei número 13.146, de 6 de julho de 2015. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência). Diário Oficial da União, Brasília, DF.

Wang, Z., Li, D., Jiang, R., and Okumura, M. (2025). Continuous Sign Language Recognition with Multi-Scale Spatial-Temporal Feature Enhancement. *IEEE Access*, 13:5491–5506.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: A Professional Framework for Multimodality Research. In *5th International Conference on Language Resources and Evaluation*, pages 1556–1559.

Yin, A., Zhong, T., Tang, L., Jin, W., Jin, T., and Zhao, Z. (2023). Gloss Attention for Gloss-Free Sign Language Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern rRecognition*, pages 2551–2562.

Yin, K. and Read, J. (2020). Better Sign Language Translation with STMC-Transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989.