

Classificação de Sinais Isolados em Libras com Descritores Cinemáticos e *Gait Energy Image*

Jonathan N. Gois^{1,2}, Miqueas dos S. Pinheiro³, Wesley L. Passos^{1,2,3}, Gabriel M. Araujo³,
Felipe da R. Henriques^{4,5,6}, Alessandro R. L. Zachi^{1,3}, Amaro A. Lima^{1,5}

Resumo—Este trabalho apresenta uma metodologia para classificação de sinais isolados em Língua Brasileira de Sinais (Libras) baseada na combinação de uma proposta de *Gait Energy Image* (GEI), características cinemáticas e filtro de Kalman. A abordagem utiliza marcos corporais e manuais normalizados, suavizados temporalmente e convertidos em descritores espaciais e dinâmicos. São avaliadas representações baseadas em GEI, características cinemáticas e sua combinação, utilizando classificadores supervisionados. Nos experimentos com a base MINDS-Libras, usando duas estratégias distintas de validação, os melhores resultados alcançaram até 98,91% de acurácia, superando referências da literatura.

Palavras-Chave—Libras, reconhecimento de sinais, descritores cinemáticos, filtro de Kalman, classificação de sinais isolados.

Abstract—This paper presents a methodology for isolated Brazilian Sign Language (Libras) recognition based on the combination of a proposed *Gait Energy Image* (GEI), kinematic features, and a Kalman filter. The proposed approach uses normalized body and hand landmarks, temporally smoothed and converted into spatial and dynamic descriptors. GEI-based features, kinematic features, and their combination are evaluated using supervised classifiers. Experiments on the MINDS-Libras dataset, using two different validation protocols, show that the best configurations achieve accuracies up to 98.91%, outperforming reference results in the literature.

Keywords—Brazilian Sign Language, sign language recognition, kinematic descriptors, Kalman filter, isolated sign classification.

I. INTRODUÇÃO

Libras é basilar à comunicação, à identidade linguística e à inclusão social dos brasileiros com deficiência auditiva [1]. Ainda assim, a comunicação entre usuários de Libras e pessoas não fluentes nessa língua permanece com limitações linguísticas e tecnológicas. Neste contexto, sistemas automáticos de reconhecimento de sinais podem apoiar o desenvolvimento de tecnologias assistivas, ferramentas educacionais e recursos de mediação comunicacional [1]–[3].

O reconhecimento automático de sinais é um desafio porque línguas de sinais combinam componentes manuais e não manuais, como configuração das mãos, orientação, posição espacial, movimento, expressões faciais e postura corporal [2]. Mesmo em sinais isolados, nos quais cada amostra contém uma unidade lexical com início e fim definidos, a classificação não é trivial. Sinais distintos podem compartilhar configurações de mão, trajetórias ou regiões de articulação semelhantes, enquanto a execução de um mesmo sinal pode variar entre sinalizadores

em amplitude, velocidade, estilo motor e posicionamento corporal [3], [4].

Métodos baseados em vídeo são promissores para aplicações de reconhecimento e tradução automática por serem mais acessíveis e pouco invasivos. Entretanto, ainda trazem consigo problemas como ruído visual, oclusões, variações de iluminação, diferenças de fundo e dependência da qualidade da câmera [2]. Com o avanço do aprendizado profundo, redes convolucionais, recorrentes e modelos com mecanismos de atenção passaram a obter resultados expressivos em diferentes bases de sinais [3]–[5]. No entanto, essas abordagens geralmente dependem de grande volume de dados anotados, treinamento custoso e maior capacidade computacional [3], [4]. Para Libras, a limitação da anotação é ainda mais saliente devido à menor disponibilidade de bases amplas, padronizadas e diversas, além de lacunas em protocolos experimentais, métricas e generalização entre usuários [1].

Tendo em vista a escassez de dados e a complexidade do sinal, torna-se interessante investigar representações compactas dos sinais a serem classificados. A técnica conhecida como *Gait Energy Image* (GEI), proposta originalmente para reconhecimento de indivíduos pela marcha, torna compacta uma sequência temporal de silhuetas em uma única imagem média, preservando a distribuição espacial recorrente do movimento [6]. Em trabalho anterior, adaptamos esse conceito para o reconhecimento de sinais em Libras, explorando representações compactas baseadas em energia em bases como CEFET/RJ-Libras [7], MINDS-Libras [8] e LIBRAS-UFOP [9]. Embora eficiente, a GEI tende a perder informação temporal fina, pois a agregação dos quadros reduz diferenças de velocidade, aceleração e variação angular ao longo do gesto.

De modo a contornar esta limitação, o presente trabalho utiliza estatísticas cinemáticas extraídas das trajetórias dos marcos corporais e manuais (*landmarks*). Esses atributos descrevem a dinâmica do movimento por meio de posição, velocidade, aceleração e componentes espectrais [10]. Como estimadores de pose podem apresentar ruídos, oscilações e falhas pontuais de rastreamento, propõe-se aqui a aplicação do filtro de Kalman às trajetórias antes da extração das características. A filtragem reduz instabilidades quadro a quadro sem eliminar a dinâmica principal do sinal [11], [12].

Em suma, as principais contribuições deste trabalho são:

- a adaptação da *Gait Energy Image* para uma representação esquelética baseada em *landmarks* corporais e manuais normalizados, em contraste com abordagens baseadas em máscaras ou silhuetas;
- a incorporação de descritores cinemáticos multi-escala, incluindo posição, velocidade, aceleração e componentes espectrais das trajetórias dos marcos;

¹CEFET-RJ, Campus Nova Iguaçu-RJ; ²PESC/COPPE/UFRJ, Rio de Janeiro-RJ; ³PPEEL/CEFET-RJ, Rio de Janeiro-RJ; ⁴CEFET-RJ, Campus Petrópolis-RJ; ⁵PPGIO/CEFET-RJ, Rio de Janeiro-RJ; ⁶PPCIC/CEFET-RJ, Rio de Janeiro-RJ. E-mails: jonathan.gois@cefet-rj.br, miqueas.pinheiro@aluno.cefet-rj.br, {wesley.passos, gabriel.araujo, felipe.henriques, alessandro.zachi, amaro.lima}@cefet-rj.br.

- o uso de filtragem de Kalman para suavização temporal das trajetórias antes da geração da GEI esquelética e da extração dos descritores cinemáticos;
- uma avaliação experimental em dois protocolos de validação sobre a base MINDS-Libras, comparando o uso isolado e combinado de GEI esquelética e características cinemáticas com resultados reportados na literatura.

O restante deste artigo está organizado da seguinte forma. A Seção II descreve a metodologia proposta, incluindo a extração dos pontos de interesse, o pré-processamento e a extração de características. A Seção III apresenta o protocolo experimental, enquanto a Seção IV discute os resultados obtidos. Por fim, a Seção V apresenta as conclusões e os trabalhos futuros.

II. METODOLOGIA

A metodologia aqui descrita é delimitada por quatro etapas: (i) extração e normalização de *landmarks*, (ii) suavização temporal por filtro de Kalman, (iii) construção da representação por GEI esquelética e (iv) extração de características cinemáticas das trajetórias suavizadas. A Fig. 1 resume o fluxo adotado, cujo objetivo é obter descritores compactos e discriminativos para a classificação de sinais isolados em Libras.

Para cada vídeo de entrada, são extraídos *landmarks* corporais e manuais utilizando a biblioteca *MediaPipe* [13]. Cada quadro é inicialmente representado por uma matriz $\mathbf{S}_t \in \mathbb{R}^{75 \times 3}$, $t = 1, \dots, T$, em que T é o número de quadros do vídeo. Os 75 pontos correspondem a 33 pontos de pose corporal e 21 pontos para cada mão, todos descritos pelas coordenadas (x, y, z) . As coordenadas x e y correspondem às posições normalizadas no plano da imagem, enquanto z representa a profundidade relativa estimada pelo *MediaPipe* [13]. Após a centralização espacial em x e y , a coordenada z é escalonada pelo mesmo fator corporal usado nas demais coordenadas, de modo a manter as três componentes em faixas comparáveis. Para as etapas de filtragem temporal, cálculo de saliência e extração dos descritores cinemáticos, essa matriz é vetorizada como $\mathbf{s}_t \in \mathbb{R}^{225}$. As coordenadas são normalizadas por sequência para mitigar variações de posição global e escala do sinalizador. Em cada quadro, os pontos são centralizados no ponto médio dos quadris, calculado a partir dos marcos 23 e 24. Em seguida, toda a sequência é escalonada pela distância média entre os ombros ao longo do vídeo. A mesma escala é aplicada às coordenadas x , y e z .

A. Representação por Gait Energy Image

A GEI condensa uma sequência temporal em uma imagem média, preservando a distribuição espacial proveniente do movimento [6]. No reconhecimento de Libras, essa ideia pode ser aplicada na dinâmica dos marcos corporais e manuais, compactando-se a informação espacial do sinal [9].

Neste trabalho, cada quadro normalizado é convertido em uma imagem binária de esqueleto $\mathbf{F}_t \in \{0, 1\}^{H \times W}$, com $H = 64$ e $W = 48$. Para isso, os marcos corporais e manuais são projetados no plano (x, y) e desenhados na imagem por meio das conexões anatômicas entre articulações adjacentes. Para evitar que cada quadro seja reescalado de forma independente, utiliza-se uma janela espacial fixa após a normalização corporal. Essa escolha preserva diferenças

relativas de amplitude e deslocamento ao longo da sequência, evitando que sinais com extensões espaciais distintas sejam artificialmente comprimidos ou expandidos para a mesma área da imagem.

Sejam $p_{t,j}^x$ e $p_{t,j}^y$ as coordenadas normalizadas do j -ésimo marco no quadro t obtidas a partir da sequência de marcos normalizada e suavizada $\mathbf{S}_t$. Considerando a faixa espacial fixa $[p_{\min}, p_{\max}]$, com $p_{\min} = -4$ e $p_{\max} = 4$, as coordenadas de pixel associadas ao marco projetado de cada coluna e linha são dadas, respectivamente, por

$$\begin{aligned} c_{t,j} &= \left\lfloor (W - 1) \frac{\text{clip}(p_{t,j}^x, p_{\min}, p_{\max}) - p_{\min}}{p_{\max} - p_{\min}} \right\rfloor, \\ r_{t,j} &= \left\lfloor (H - 1) \left(1 - \frac{\text{clip}(p_{t,j}^y, p_{\min}, p_{\max}) - p_{\min}}{p_{\max} - p_{\min}} \right) \right\rfloor, \end{aligned} \quad (1)$$

onde $\text{clip}(\cdot)$ refere-se à operação de limitação do valor entre os limites pré-estabelecidos.

A GEI uniforme é obtida pela média temporal dos quadros renderizados:

$$\mathbf{G}_{\text{unif}} = \frac{1}{T} \sum_{t=1}^T \mathbf{F}_t. \quad (2)$$

Embora essa representação sintetize as regiões mais frequentemente ocupadas pelo corpo e pelas mãos, ela atribui o mesmo peso a trechos estáticos e a instantes de maior movimento. Para enfatizar quadros mais informativos, utiliza-se uma versão ponderada por saliência temporal. Primeiro, define-se

$$d_t = \|\mathbf{s}_t - \mathbf{s}_{t-1}\|_2, \quad t = 2, \dots, T, \quad (3)$$

e os pesos são calculados por

$$\omega_t = \frac{d_t}{\sum_{\tau=2}^T d_{\tau} + \epsilon}, \quad t = 2, \dots, T; \quad \epsilon > 0. \quad (4)$$

Para o primeiro quadro, adota-se $\omega_1 = \omega_2$. A GEI ponderada por saliência é então definida como

$$\mathbf{G}_{\text{sal}} = \frac{\sum_{t=1}^T \omega_t \mathbf{F}_t}{\sum_{t=1}^T \omega_t + \epsilon}. \quad (5)$$

Por fim, tem-se $\mathbf{G}_{\text{sal}}$ vetorizada como um descritor de dimensão $H \times W = 3072$. Neste trabalho, usa-se isoladamente ou concatenado às características cinemáticas na etapa de classificação. A principal diferença em relação a abordagens anteriores baseadas em GEI [9] está na construção da representação: em vez de calcular a imagem de energia a partir de máscaras densas de partes corporais, este trabalho renderiza esqueletos obtidos de marcos anatômicos normalizados e suavizados, preservando a geometria relativa do movimento.

B. Extração de Características Cinemáticas

Além da representação espacial por GEI, são extraídas características cinemáticas das trajetórias normalizadas e suavizadas pelo filtro de Kalman. Esses descritores representam a dinâmica do sinal por meio de posição, velocidade, aceleração e componentes espectrais, uma vez que sinais distintos podem depender não apenas da configuração das mãos, mas também da direção, amplitude, velocidade e frequência do movimento [3], [4], [10].

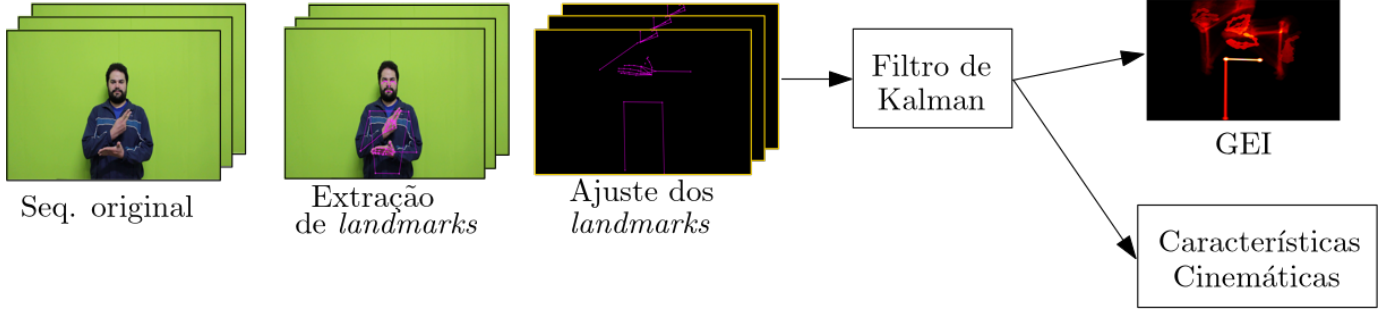


Fig. 1. Fluxo da metodologia proposta, incluindo extração e ajuste dos *landmarks*, suavização por filtro de Kalman, geração da GEI e extração de características cinemáticas.

Seja $\mathbf{s}_t \in \mathbb{R}^{225}$ o vetor de coordenadas dos 75 marcos tridimensionais no instante t , após normalização espacial e suavização temporal. A componente de posição é dada, simplesmente, por $\mathbf{p}_t = \mathbf{s}_t$. As velocidades são calculadas por diferenças em múltiplas escalas temporais:

$$\mathbf{v}_t^{(\delta)} = \mathbf{s}_t - \mathbf{s}_{t-\delta}, \quad \delta \in \{1, 2, 4\}. \quad (6)$$

Essa escolha permite representar deslocamentos rápidos, observáveis entre quadros adjacentes, e movimentos mais amplos, evidenciados em intervalos temporais maiores. A aceleração é obtida por diferença de segunda ordem:

$$\mathbf{a}_t = \mathbf{s}_{t+1} - 2\mathbf{s}_t + \mathbf{s}_{t-1}. \quad (7)$$

Nas bordas da sequência, são usadas replicações dos quadros extremos para manter a dimensão dos descritores. Como velocidades e acelerações dependem de diferenças temporais, a suavização prévia reduz a amplificação de ruídos de rastreamento.

As características espectrais das trajetórias também são extraídas neste passo. Para uma sequência de comprimento T , considera-se uma janela temporal de comprimento

$$L_{\text{fft}} = \min \left(\max \left(\left\lfloor \frac{T}{\alpha} \right\rfloor, L_{\min} \right), L_{\max} \right), \quad (8)$$

com $\alpha = 4$, $L_{\min} = 16$ e $L_{\max} = 64$, cujos parâmetros são definidos conforme [10]. A sequência é preenchida nas bordas pela replicação do primeiro e do último quadro, garantindo janelas de mesmo comprimento ao longo de todo o vídeo. Para cada coordenada, aplica-se uma janela de Hann e calcula-se a Transformada Rápida de Fourier na vizinhança temporal centrada em t . Define-se uma faixa de frequências para seleção da frequência dominante $f \in [0, 3; 5, 0]$ Hz, armazenando-se sua magnitude e sua fase normalizada:

$$\mathbf{m}_t \in \mathbb{R}^{225}, \quad \phi_t \in \mathbb{R}^{225}. \quad (9)$$

O descritor cinemático por quadro é formado pela concatenação das componentes de posição, velocidades multi-escala, aceleração, magnitude espectral e fase:

$$\mathbf{f}_t = [\mathbf{p}_t; \mathbf{v}_t^{(1)}; \mathbf{v}_t^{(2)}; \mathbf{v}_t^{(4)}; \mathbf{a}_t; \mathbf{m}_t; \phi_t] \in \mathbb{R}^{1575}. \quad (10)$$

Como cada componente possui 225 valores, obtêm-se $7 \times 225 = 1575$ atributos por quadro.

Para obter uma representação de dimensão fixa, a sequência $\{\mathbf{f}_t\}_{t=1}^T$ é agregada por estatísticas globais calculadas atributo a atributo:

$$\mathbf{k} = [\boldsymbol{\mu}; \boldsymbol{\sigma}; \mathbf{f}_{\max}; \mathbf{f}_{\min}] \in \mathbb{R}^{6300}. \quad (11)$$

Assim, a matriz temporal de dimensão $T \times 1575$ é convertida em um vetor de $4 \times 1575 = 6300$ atributos. O vetor cinemático pode ser usado isoladamente ou concatenado ao descritor GEI. Com o intuito de reduzir a possibilidade de sobreajuste dos métodos de classificação, aplica-se a SVD truncada para redução de dimensionalidade, seguida de padronização dos atributos.

C. Filtro de Kalman

As trajetórias obtidas por estimadores de pose podem apresentar ruídos, oscilações e falhas pontuais de rastreamento, sobretudo em casos de oclusão, movimentos rápidos, baixa resolução ou variações de iluminação. Esses efeitos prejudicam a extração de características cinemáticas, pois velocidades e acelerações são calculadas por diferenças temporais e tendem a amplificar variações quadro a quadro. Para reduzir esse problema, aplica-se um filtro de Kalman às sequências de marcos antes da construção da GEI e da extração dos descritores cinemáticos.

Cada coordenada dos marcos é filtrada de forma independente por um modelo escalar de caminhada aleatória. Seja y_t a coordenada observada no instante t e x_t o estado suavizado correspondente. O modelo adotado é

$$x_t = x_{t-1} + w_t, \quad y_t = x_t + \nu_t, \quad (12)$$

com $w_t \sim \mathcal{N}(0, Q)$ e $\nu_t \sim \mathcal{N}(0, R)$. Neste trabalho, adotam-se $Q = 10^{-4}$ e $R = 5 \times 10^{-3}$, o que resulta em uma suavização moderada das trajetórias. As etapas de predição e atualização seguem o filtro de Kalman escalar convencional.

O filtro é aplicado a todas as coordenadas da sequência normalizada $\mathbf{s}_t \in \mathbb{R}^{225}$, correspondente aos 75 marcos tridimensionais de pose corporal e mãos. Como a filtragem é realizada em uma única passagem temporal, o procedimento é causal e não utiliza quadros futuros. A inicialização é feita a partir do primeiro quadro válido da sequência, definido por

$$t_0 = \min \{t : \|\mathbf{s}_t\|_2 > 10^{-6}\}. \quad (13)$$

Essa escolha evita iniciar o filtro com um vetor nulo quando há falhas de detecção nos primeiros quadros. A sequência suavizada resultante é usada tanto na renderização da GEI quanto na extração das características cinemáticas.

III. EXPERIMENTO

Os experimentos avaliam a capacidade discriminativa das representações propostas na classificação de sinais isolados em Libras. Foram consideradas três configurações de entrada: GEI, usando apenas a representação por *Gait Energy Image*; KIN,

usando apenas características cinemáticas; e COMB, obtida pela concatenação dos dois descritores. Essa comparação permite analisar a contribuição individual e combinada da representação espacial compacta e dos atributos dinâmicos do movimento.

A base utilizada foi a MINDS-Libras, composta originalmente por sinais isolados distribuídos em 20 classes lexicais, executados por 12 sinalizadores, com cinco repetições por sinal [8]. Cada sequência corresponde a uma ocorrência de um sinal isolado com início e fim definidos. Para cada vídeo, os marcos corporais e manuais foram extraídos, normalizados, suavizados por filtro de Kalman e convertidos nas representações descritas na Seção II. A GEI foi obtida pela renderização dos esqueletos ao longo do tempo, enquanto o vetor cinemático reuniu posição, velocidades multi-escala, aceleração e componentes espectrais das trajetórias.

Foram avaliados dois protocolos experimentais. No Prot-A, utilizou-se uma divisão aleatória estratificada com 75% das amostras para treino e 25% para teste, repetida em 10 partições independentes, seguindo a configuração de comparação com Passos *et al.* [9]. No Prot-B, utilizou-se validação cruzada estratificada com 10 *folds*, permitindo comparação indicativa com Rego *et al.* [10]. Em ambos os protocolos, a divisão foi realizada por amostra, e não por sinalizador; portanto, vídeos de um mesmo sinalizador podem aparecer em treino e teste em partições distintas. Em ambos os casos, os resultados são apresentados em termos de acurácia média e desvio padrão.

Em todos os protocolos, as etapas de redução de dimensionalidade por SVD, padronização e treinamento dos classificadores foram ajustadas exclusivamente nos dados de treino de cada *split/fold*, sendo posteriormente aplicadas ao conjunto de teste correspondente.

A classificação foi realizada com quatro classificadores supervisionados: máquina de vetores de suporte com núcleo RBF (SVM), perceptron multicamadas (MLP), *Extra Trees* (ET) e *Random Forest* (RF). Antes do treinamento, os vetores de características foram submetidos à redução de dimensionalidade por SVD truncada e à padronização dos atributos.

IV. RESULTADOS E DISCUSSÃO

A Fig. 2 apresenta a distribuição dos resultados obtidos pelos métodos avaliados nos protocolos considerados. Observa-se que os métodos baseados apenas em GEI apresentam desempenho superior ao resultado de referência de Passos *et al.* [9] em algumas configurações, especialmente para o classificador SVM. No Prot-A, o método GEI+SVM alcançou acurácia de $89,62\% \pm 1,74\%$, superando os $84,66\%$ reportados por Passos *et al.* [9]. Esse resultado indica que a representação GEI, quando construída a partir de uma normalização espacial consistente e combinada a um classificador adequado, ainda fornece uma descrição compacta e relevante para sinais isolados em Libras.

Apesar do bom desempenho da GEI, os melhores resultados foram obtidos pelos métodos baseados em características cinemáticas e pela combinação entre descritores cinemáticos e GEI. No Prot-A, o melhor resultado foi alcançado pelo método KIN+MLP, com acurácia de $98,90\% \pm 0,69\%$. Esse desempenho sugere que as informações de posição, velocidade, aceleração e frequência extraídas das trajetórias dos marcos são altamente discriminativas para a classificação de sinais isolados.

TABELA I
COMPARAÇÃO DE DESEMPENHO ENTRE OS MÉTODOS AVALIADOS PARA PROT-A E PROT-B.

Método	Prot-A (%)	Prot-B (%)
Passos <i>et al.</i> (2021) [9]	$84,66 \pm 1,78$	–
Rego <i>et al.</i> (2025) [10]	–	$92,00 \pm 3,00$
GEI+SVM	$89,62 \pm 1,74$	$92,82 \pm 1,14$
GEI+MLP	$87,74 \pm 2,32$	$89,81 \pm 2,25$
GEI+ET	$83,77 \pm 1,39$	$86,99 \pm 2,77$
KIN+SVM	$96,72 \pm 1,04$	$97,76 \pm 1,35$
KIN+MLP	$98,90 \pm 0,69$	$98,40 \pm 1,04$
KIN+ET	$97,23 \pm 0,74$	$97,44 \pm 1,40$
COMB+SVM	$95,62 \pm 1,58$	$97,95 \pm 1,56$
COMB+MLP	$98,67 \pm 0,66$	$98,91 \pm 0,76$
COMB+ET	$96,92 \pm 0,99$	$97,56 \pm 1,40$
COMB+RF	$95,41 \pm 1,11$	$97,69 \pm 0,92$

No Prot-B, o melhor resultado foi obtido pelo método COMB+MLP, com acurácia de $98,91\% \pm 0,76\%$. Esse desempenho é superior ao resultado reportado por Rego *et al.* [10] para a base MINDS-Libras em validação cruzada 10-*fold*, no qual os autores obtiveram $92,00\% \pm 3,00\%$ de acurácia. Como os detalhes de particionamento, aumento de dados e treinamento não são necessariamente idênticos entre os trabalhos, essa comparação deve ser interpretada como indicativa sob protocolos aproximados, e não como uma reprodução estritamente controlada dos experimentos originais.

A Tabela I resume os resultados obtidos. Nota-se que os métodos KIN+SVM, KIN+MLP, KIN+ET, COMB+SVM, COMB+MLP, COMB+ET e COMB+RF apresentam acurácias superiores a 95% em pelo menos um dos protocolos avaliados. Esse comportamento reforça a relevância dos descritores cinemáticos para o problema, uma vez que os métodos baseados em KIN superaram consistentemente as versões baseadas apenas em GEI.

Comparando os classificadores, a MLP apresentou os melhores resultados gerais. Em particular, KIN+MLP foi o melhor método no Prot-A e COMB+MLP foi o melhor no Prot-B. Os métodos baseados em árvores, embora competitivos, apresentaram desempenho ligeiramente inferior aos melhores resultados com MLP e SVM. Já os classificadores SVM demonstraram desempenho elevado quando combinados às características cinemáticas, alcançando $96,72\% \pm 1,04\%$ no Prot-A e $97,76\% \pm 1,35\%$ no Prot-B.

Os resultados também mostram que a concatenação entre GEI e características cinemáticas nem sempre supera o uso isolado das características cinemáticas. No Prot-A, por exemplo, KIN+MLP obteve $98,90\%$, enquanto COMB+MLP obteve $98,67\%$. No Prot-B, entretanto, a combinação foi ligeiramente superior, alcançando $98,91\%$ com COMB+MLP. Esse comportamento sugere que a GEI pode fornecer informação complementar em determinados cenários, mas que a maior parte da discriminação entre os sinais avaliados está concentrada nos descritores cinemáticos.

De forma geral, os resultados indicam que a metodologia proposta é capaz de superar as referências consideradas nos protocolos avaliados. Em relação a Passos *et al.*, o melhor resultado no Prot-A representa um ganho absoluto de 14,24 pontos percentuais. Em relação a Rego *et al.*, o melhor resultado no Prot-B representa um ganho absoluto de 6,91 pontos percentuais. Esses ganhos reforçam a hipótese de que a

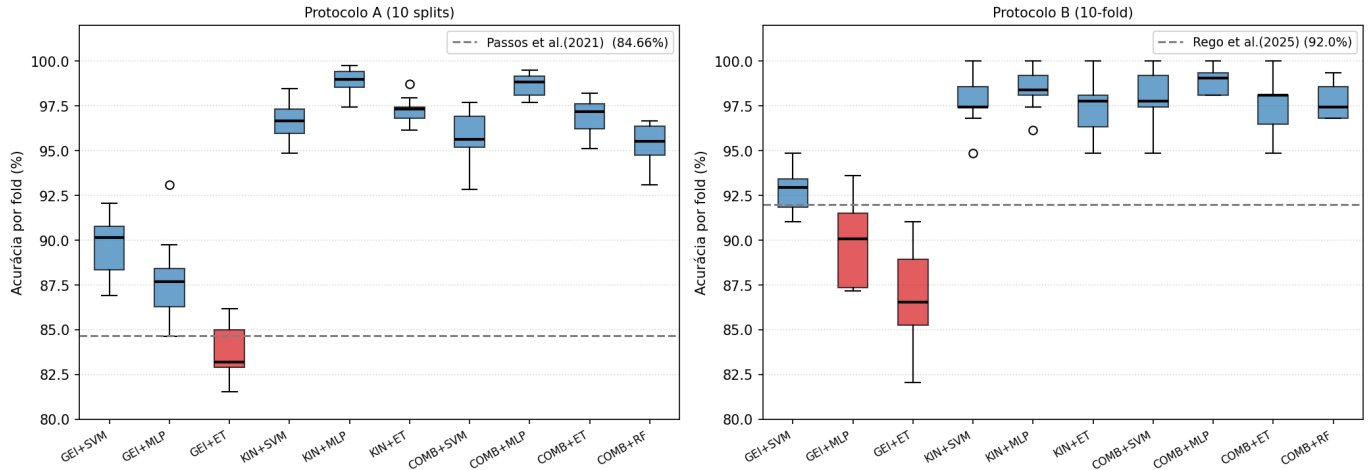


Fig. 2. Distribuição das acurácias obtidas pelos métodos avaliados nos protocolos experimentais considerados.

combinação entre filtragem temporal, descritores cinemáticos multi-escala e representações compactas por GEI é adequada para a classificação de sinais isolados em Libras.

Embora os resultados obtidos sejam elevados, os protocolos avaliados não necessariamente medem a generalização para sinalizadores não vistos. Assim, os ganhos reportados devem ser entendidos no contexto da classificação de sinais isolados sob os particionamentos avaliados, sendo a validação *cross-subject* uma etapa importante para trabalhos futuros.

V. CONCLUSÕES

Este trabalho apresentou uma metodologia para classificação de sinais isolados em Libras baseada na combinação de *Gait Energy Image*, características cinemáticas e filtro de Kalman. A abordagem proposta busca explorar informações complementares de forma e movimento, mantendo uma representação compacta e compatível com classificadores supervisionados tradicionais.

Os resultados experimentais indicaram que as características cinemáticas extraídas das trajetórias dos marcos são altamente discriminativas para o problema avaliado. O método KIN+MLP obteve o melhor desempenho no Prot-A, com $98,90\% \pm 0,69\%$, enquanto o método COMB+MLP obteve o melhor desempenho no Prot-B, com $98,91\% \pm 0,76\%$. Ambos os resultados superaram as referências consideradas na literatura para os protocolos analisados [9], [10].

A análise comparativa mostrou que a GEI fornece uma representação espacial compacta e relevante, mas que seu uso isolado tende a ser menos discriminativo do que os descritores cinemáticos. Por outro lado, a combinação entre GEI e características cinemáticas pode fornecer ganhos em determinados protocolos, sugerindo que as duas representações carregam informações parcialmente complementares.

Como trabalhos futuros, pretende-se avaliar a metodologia em protocolos mais desafiadores de generalização entre sinalizadores, investigar estratégias de fusão mais seletivas entre GEI e características cinemáticas e analisar o impacto individual de cada grupo de descritores. Também se pretende explorar arquiteturas temporais capazes de preservar a ordem completa dos quadros, de modo a comparar representações estatísticas compactas com modelos sequenciais capazes de preservar explicitamente a ordem temporal dos quadros.

AGRADECIMENTOS

Os autores agradecem o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), por meio da Chamada Universal, processo 403327/2025-3; do Projeto Brasil 6G RNP/MCTI, processo 01245.020548/2021-07; à Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) - Processo E-26/210.425/2024 e dos editais internos de fomento à pesquisa do CEFET/RJ.

REFERÊNCIAS

- [1] V. Costa, J. P. Tomaszewski, L. Pereira, B. S. Santana, and G. Corrêa, "AI-based approaches for Brazilian sign language recognition: A systematic literature review: Insights on methods, metrics, and resources for Libras recognition," in *Proc. Braz. Symp. Multim. Web*, 2025, pp. 585–597.
- [2] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 1, pp. 131–153, 2019.
- [3] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Syst. Appl.*, vol. 164, p. 113794, 2021.
- [4] N. Sarhan and S. Frintrop, "Unraveling a decade: A comprehensive survey on isolated sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3210–3219.
- [5] A. Wadhawan and P. Kumar, "Deep learning-based sign language recognition system for static signs," *Neural Comput. Appl.*, vol. 32, no. 12, pp. 7957–7968, 2020.
- [6] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, 2006.
- [7] P. V. Gameiro, W. L. Passos, G. M. Araujo, A. A. Lima, J. N. Gois, and A. R. Corbo, "A Brazilian sign language video database for automatic recognition," in *Proc. IEEE Lat. Am. Rob. Symp.*, 2020, pp. 1–6.
- [8] T. Rezende, S. Almeida, and F. Guimarães, "Development and validation of a Brazilian sign language database for human gesture recognition," *Neural Comput. Appl.*, vol. 1, pp. 1–19, Mar. 2021.
- [9] W. L. Passos, G. M. Araujo, J. N. Gois, and A. A. de Lima, "A gait energy image-based system for Brazilian sign language recognition," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 68, no. 11, pp. 4761–4771, 2021.
- [10] R. C. Rego, L. M. de Moraes, and W. M. Almeida, "Brazilian sign language recognition using deep learning based on fast fourier transform and kinematic features," *IEEE Access*, vol. 13, pp. 202 875–202 892, 2025.
- [11] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.
- [12] B. Liu, S. Gao, Y. Wang, and Y. Wei, "Dynamic kalman fusion for robust continuous sign language recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2026, pp. 19 122–19 126.
- [13] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Ubaweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee *et al.*, "Mediapipe: A framework for perceiving and perceiving systems," in *Proc. Int. Conf. Mach. Learn. Workshop Comput. Vis. AR/VR*, 2019.