4SkeletonFeat: A Framework for Unifying the Extraction of Features from Videos of Human Activities

Diogo H. Godoi[1], Roseli A. F. Romero[1],Caetano Mazzoni Ranieri[2], Patricia A. Vargas[3],

[1] Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP – São Carlos)
Caixa Postal 15.064 – 13566-590 – São Carlos – SP – Brazil

[2]Instituto de Geociências e Ciências Exatas
Universidade Estadual Paulista (UNESP – Rio Claro)
Av. 24 A, 1515 - Bela Vista, Rio Claro - SP, 13506-700

[3]Edinburgh Centre for Robotics
Heriot-Watt University (Edimburgo - UK)
Riccarton Campus, Edimburgo EH14 4AS, Reino Unido

diogo.godoi@usp.br, rafrance@icmc.usp.br, cm.ranieri@unesp.br, P.A.Vargas@hw.ac.uk

Abstract. *This article describes a framework for unifying feature extraction from videos of human activities. The proposed system, 4SkeletonFeat, provides a set of tools and libraries that allow researchers to easily extract Joint, Bone, Bone-Motion, and Joint-Motion features. It also offers tools for flexible feature representation, including token-based, graph-based, pseudo-image-based, and sequence-based formats. To evaluate the framework, a new dataset called FitTube was developed from YouTube videos. Results show that 4SkeletonFeat accurately and efficiently extracts these features, providing effective representations for machine learning tasks.*

Resumo. *Neste artigo, é proposto um framework para unificar a extração de características de vídeos de atividades humanas. O framework, chamado 4SkeletonFeat, fornece um conjunto de ferramentas e bibliotecas que permitem aos pesquisadores extrair facilmente características de vídeos de atividades humanas. Ele permite extrair características como: "Joint", "Bone", "Bone-Motion"e "Joint-Motion"de vídeos de atividades humanas. Este framework também fornece um conjunto de ferramentas para representação dessas características, como representações baseadas em tokens, representações baseadas em grafos, representações pseudo-imagem e representações baseadas em sequência. Para avaliar o framework proposto, um novo conjunto de dados chamado FitTube foi desenvolvido a partir de vídeos do Youtube. Os resultados mostram que 4SkeletonFeat é capaz de extrair características de vídeos de atividades humanas e fornecer uma boa representação para essas características para tarefas de Aprendizado de Máquina.*

1. Introdução

Com o avanço dos estudos na área de visão computacional, a extração de características de vídeos de atividades humanas tem se tornado cada vez mais importante para tarefas como reconhecimento de atividades, análise de comportamento e monitoramento de saúde.

No trabalho de Liu et al. [Liu et al. 2025], é apresentada uma análise de cada uma das técnicas existentes que utilizam o esqueleto 3D para extrair características. É destacado que o uso do esqueleto 3D já oferece a topologia e a estrutura do corpo humano através de suas articulações e ossos, o que é uma vantagem para a extração de características. Além disso, o esqueleto 3D é mais robusto a variações de iluminação e oclusões em comparação com outras técnicas de extração de características.

Apesar das vantagens inerentes ao uso de esqueletos 3D, o processamento de vídeo bruto (RGB) enfrenta desafios críticos, como a variabilidade visual decorrente de mudanças de iluminação, fundos complexos e oclusões, que corrompem diretamente o sinal observado [Alomar et al. 2024, Xie et al. 2025]. O processamento de vídeos completos exige alto custo computacional, de armazenamento e de treinamento, frequentemente demandando o uso de CNNs 3D ou Transformers [Abdelkawy et al. 2024]. Por outro lado, embora os dados de esqueleto removam redundâncias ambientais, a extração dessas poses por modelos como MediaPipe ou OpenPose constitui um gargalo significativo devido à latência de inferência e à complexidade dos pipelines de múltiplos estágios [Gao et al. 2025, Rode et al. 2025, Neupane et al. 2024].

No campo do reconhecimento de atividades humanas (HAR), a literatura aponta uma evolução das arquiteturas de redes neurais recorrentes (RNN) e convolucionais (CNN) para modelos baseados em grafos (GCN) e Transformers [Pellano et al. 2024, Tu et al. 2025, Xiang and Wang 2024]. O estado da arte atual é definido pela capacidade de modelagem de topologias adaptativas e pelo raciocínio híbrido global-local [Tu et al. 2025]. Nesse contexto, a representação multimodal — que combina juntas, ossos e informações de movimento (motion) — tem se mostrado superior a abordagens unimodais, pois cada fluxo captura aspectos complementares da geometria e da dinâmica temporal do corpo humano [Zhang et al. 2024, Halmich et al. 2025]. A fusão dessas características cinemáticas é essencial para aumentar a precisão e a robustez do reconhecimento em cenários do mundo real, onde ruídos e variações de ponto de vista são comuns.

Para resolver os problemas apontados, neste trabalho está sendo proposto o sistema *4SkeletonFeat*[1], que é um *framework* unificado para a extração de características de vídeos de atividades humanas. Este *framework* fornece um conjunto de ferramentas e bibliotecas que permitem aos pesquisadores extrair facilmente características como as articulações (”Joint”), ossos (”Bone”), movimento dos ossos (”Bone-Motion”) e movimento das articulações (”Joint-Motion”) de vídeos de atividades humanas. Além disso, o *framework 4SkeletonFeat* também oferece ferramentas para representar essas características em diferentes formatos, como representações baseadas em tokens, grafos e pseudo-imagens. Com isso, para avaliar o *4SkeletonFeat*, foi desenvolvida uma nova base de dados denominada FitTube, que consiste em vídeos do Youtube relacionados a atividades físicas. Os resultados demonstram que o *4SkeletonFeat* é capaz de extrair características, fornecendo uma representação eficaz para tarefas de Aprendizado de Máquina.

[1] `https://github.com/diogodoi/4SkeletonFeat`

2. Trabalhos Relacionados

No contexto atual do Reconhecimento de Ações Humanas baseado em Esqueleto (*Skeleton-Based Action Recognition* – HAR), a literatura aponta que frameworks recentes para essa tarefa vêm unificando a extração espacial e temporal por meio de arquiteturas paralelas do tipo GCN-Transformer, modelos de fundação auto-supervisionados e estratégias de codificação de características multimodais [Chen et al. 2025, Wang et al. 2026, Xie et al. 2025]. *Frameworks* utilitários focados na engenharia unificada das *features* extraídas como o *4SkeletonFeat* posiciona-se como uma nova categoria, diferenciando-se fundamentalmente das arquiteturas dominantes recentes como USDRL [Wang et al. 2026], SA-TDGFormer [Chen et al. 2025] e SkelFormer [Yan et al. 2026]. Esta seção detalha as principais distinções técnicas que validam a relevância do *framework* proposto.

Os *frameworks* mais recentes citados (e.g., USDRL, SA-TDGFormer) geralmente atuam como modelos completos de reconhecimento, onde o treinamento ocorre *end-to-end* com foco em otimizar pesos específicos para tarefas *downstream* definidas (como detecção ou classificação). Em contrapartida, o *4SkeletonFeat* atua na infraestrutura de pré-processamento e extração. Enquanto métodos recentes integram diretamente redes neurais convolucionais em grafos (GCN) ou *Transformers* nos fluxos principais, com modelos de predição embutidos para maximizar a acurácia [Wang et al. 2026], o *4SkeletonFeat* oferece uma camada modular que permite extrair características semânticas independentemente da arquitetura final de inferência.

Isso confere ao *framework* maior flexibilidade: ele não força um paradigma único de modelagem, mas sim fornece representações versáteis (tokens discretos, pseudo-sequências ou grafos esparsos) que podem ser injetadas em modelos clássicos, redes recorrentes (LSTM, RNN) e arquiteturas modernas como *Transformers*. A análise comparativa sugere que a escolha do *4SkeletonFeat* é preferencial quando o objetivo não é criar um classificador específico desde a base (como no caso do USDRL), mas sim padronizar os *inputs* para diferentes *pipelines* de *Deep Learning* existentes ou futuros desenvolvimentos comunitários.

Em síntese, o *4SkeletonFeat* diferencia-se dos *frameworks* recentes (2025 - 2026) não pela sua capacidade preditiva em tarefas específicas, que é garantida por modelos como USDRL e SA-TDGFormer, mas pelo seu papel como ferramenta de *feature representation* unificada. A proposta oferece uma infraestrutura padrão para a extração cinemática, permitindo múltiplas formas de representações (tokens, grafos), que não são abordados por *pipelines end-to-end* convencionais focados apenas na modelagem do reconhecimento.

3. Materiais e Métodos

Para o desenvolvimento do framework *4SkeletonFeat*, foi adotada uma abordagem modular e escalável, estruturada em componentes independentes especializados na extração e na engenharia de características biomecânicas. Implementado em linguagem Python, o ecossistema integra ferramentas como OpenCV, Matplotlib, NetworkX, NumPy e PyTorch para viabilizar processamentos eficientes de fluxos cinemáticos. O pipeline foi projetado para operar de forma transparente em relação à fonte de dados originais, sendo compatível com saídas de sensores de profundidade (e.g., Kinect) ou *frameworks* de esti-

mativa de pose baseados em visão computacional como MediaPipe, YOLO-Pose e OpenPose.

O fluxo de processamento inicia-se com a padronização dos dados espaciais brutos na pasta de entrada através do Estimador de Pose Espacial Primitiva (EPEP), conforme é mostrado em na Figura 1 em a). Independentemente do sensor de captura, as poses são estruturadas em arquivos NumPy comprimidos (`.npz`) sob uma matriz tridimensional com dimensões correspondendo a $(Frames, Número\ de\ Nós, 4)$. Os eixos representam, respectivamente, a linha temporal, as articulações biológicas mapeadas e um vetor de atributos contendo as coordenadas espaciais absolutas, organizadas estritamente na ordem $[X, Y, Z, \text{Visibilidade}]$, onde a visibilidade constitui uma métrica de confiança normalizada no intervalo $[0.0, 1.0]$.

Para viabilizar o cálculo correto das relações cinemáticas, o usuário deve configurar previamente o modelo anatômico no Registro de Topologia, especificando o número total de nós (N) e a lista de arestas conexas ($\mathcal{E}$) que definem os segmentos ósseos do esqueleto de interesse. A partir dessa parametrização, o *framework* executa, sequencialmente, os extratores de características básicas (Pose Bruta e Vetores de Ossos), gerando representações invariantes à translação global do indivíduo através do cálculo de vetores direcionais anatômicos.

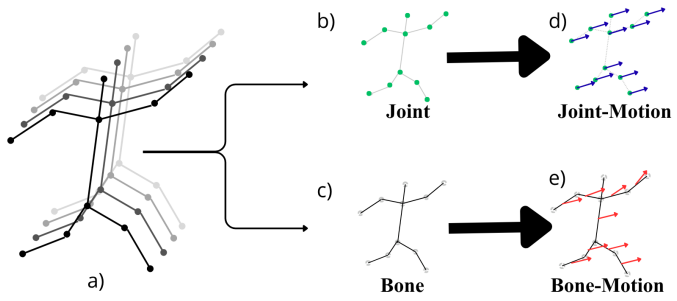


Figura 1. Representação esquemática do fluxo de trabalho do *framework 4SkeletonFeat*. O *pipeline* inicia-se com a padronização das poses brutas, seguida pela extração de características cinemáticas.

As propriedades cinemáticas e a velocidade do movimento humano são computadas de forma diferencial entre quadros (*frames*) consecutivos por meio dos módulos Descritor Cinemático de Articulações (DCA) e Descritor Cinemático de Segmentos Ósseos (DCO). Com base no princípio cinemático de diferenças finitas discretas discutido por Liu et al. [Liu et al. 2025], o *framework* computa a taxa de variação temporal de posições e orientações sob uma premissa de amostragem uniforme. Formalmente, o deslocamento diferencial para articulações e ossos em um instante t é modelado pelas equações a seguir, gerando tensores de saída com dimensões temporais decrescidas em uma unidade, isto é, $(Frames - 1, N, 4)$:

$$\text{Joint-Motion}_t = \frac{\text{Joint}_t - \text{Joint}_{t-1}}{\Delta t} \tag{1}$$

$$\text{Bone-Motion}_t = \frac{\text{Bone}_t - \text{Bone}_{t-1}}{\Delta t} \tag{2}$$

O grande diferencial do *4SkeletonFeat* reside na sua capacidade de reestruturar essas características cinemáticas em múltiplos paradigmas representacionais estruturados. Desse modo, o usuário pode exportar os mesmos dados biomecânicos em formatos nativamente otimizados para serem usados em diferentes modelos de Aprendizado de Máquina e Aprendizado Profundo, subdivididos em representações baseadas em tokens, grafos, pseudo-imagens e sequências contínuas.

3.1. Representações das Características

Como mencionado na seção anterior, o framework é capaz de produzir diversas representações para algoritmos de aprendizado de máquina. A seguir, serão apresentadas, de forma sistemática, cada uma das representações que ele é capaz de gerar.

3.1.1. Representação Baseada em Tokens Discretos para Transformers

A abordagem de tokenização do framework *4SkeletonFeat* adota a estratégia de Quantização Vetorial (*Vector Quantization*). O objetivo é converter as geometrias complexas e contínuas de um esqueleto inteiro em uma sequência discreta de símbolos linguísticos estruturados, viabilizando o processamento direto por arquiteturas baseadas em mecanismos de autoatenção (*self-attention*), como os Transformers (e.g., BERT, GPT).

O *framework* emprega o algoritmo de agrupamento K-Means (ou *MiniBatchKMeans*) para construir um dicionário de posturas universais (*Codebook*), contendo $K = 256$ centroides espaciais. O dicionário é treinado de forma estratificada entre as classes do *dataset*. Durante a inferência, a matriz contínua de um quadro no instante t é mapeada para o índice inteiro do centroide mais próximo segundo a distância euclidiana, transformando o fluxo temporal de poses em um vetor unidimensional de formato $(\text{Frames},)$. Além disso, também é possível que o usuário configure o número de centroides K e o método de agrupamento para personalizar a granularidade e a estrutura do vocabulário de tokens, permitindo um controle fino sobre o trade-off entre compressão e fidelidade da representação.

Posteriormente, ao alimentar uma rede neural profunda, esses IDs inteiros são projetados em um espaço latente denso através de uma camada de incorporação (*Embedding*), operando em conjunto com uma função de codificação posicional (*Positional Encoding*) para reinserir as noções de ordenação temporal e topológica no modelo. A representação final de um token incorporado $\mathbf{z}_t$ injetado no Transformer é dada por:

$$\mathbf{z}_t = \text{PoseEmbedding}(\mathbf{s}_t) + \text{PosEmbed}(t)$$

onde $\mathbf{s}_t \in \{0, 1, \ldots, 255\}$ é o token discreto fornecido pelo framework no frame t, PoseEmbedding traduz este índice para um vetor denso de dimensão C_{in}, e PosEmbed atribui a assinatura cronológica necessária através de funções senoidais e cossenoidais de frequências variadas, expressas por:

$$\text{PE}(p, i) = \begin{cases} \sin\left(\frac{p}{10000^{i/C_{in}}}\right) & \text{se } i \text{ é par} \\ \cos\left(\frac{p}{10000^{i/C_{in}}}\right) & \text{se } i \text{ é ímpar} \end{cases}$$

onde p correlaciona-se com o índice temporal do quadro e i delimita a dimensão interna do vetor de incorporação.

3.1.2. Representação Baseada em Grafos

Para modelos estruturais baseados em Redes de Convolução em Grafos (GCN e ST-GCN), o módulo Codificador Topológico de Grafos (CTG) traduz o esqueleto em estruturas matemáticas não-direcionadas capazes de preservar explicitamente a conectividade biológica e anatômica do corpo humano. O framework calcula uma matriz de adjacência reflexiva ($\mathbf{A} \in \mathbb{R}^{N \times N}$), que incorpora conexões funcionais automáticas (*self-loops*) em cada nó para garantir a retenção de características locais durante as operações de agregação de mensagens na rede convolucional.

A saída gerada pelo CTG é formatada para garantir compatibilidade nativa com ferramentas de aprendizado em grafos, como o *PyTorch Geometric (PyG)*. Os dados são exportados contendo os atributos dinâmicos dos nós $x \in \mathbb{R}^{Frames \times N \times 4}$ e a lista de conectividade mapeada sob o formato de coordenadas esparsas COO ($\texttt{edge_index} \in \mathbb{R}^{2 \times |\mathcal{E}|}$), conforme ilustrado na Figura 2. Essa estruturação permite o mapeamento simultâneo das adjacências espaciais físicas do corpo e de suas evoluções no domínio do tempo.

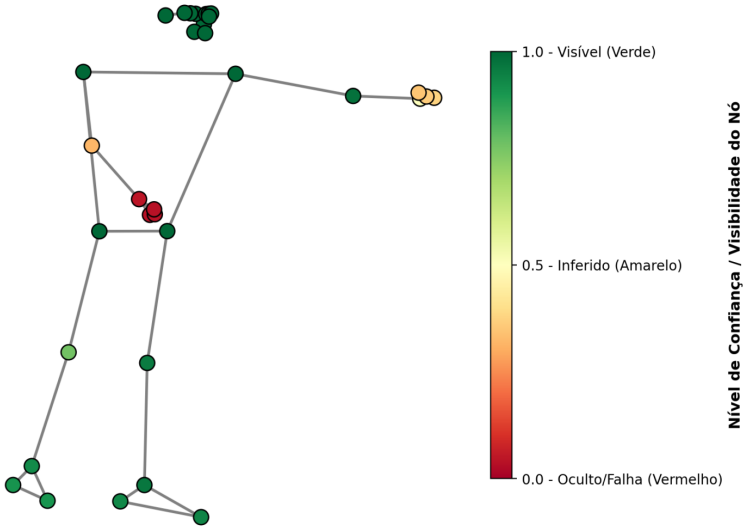


Figura 2. Exemplo de representação baseada em grafos para um esqueleto 3D. Os nós representam as articulações e os atributos cinemáticos, enquanto as arestas delimitam a conectividade anatômica e os mapeamentos estruturais.

3.1.3. Representação por Pseudo-Imagens

Com o objetivo de aproveitar a alta eficiência de arquiteturas convolucionais 2D tradicionais fortemente consolidadas na visão computacional (e.g., ResNet, MobileNet, Effici-

entNet), o módulo Codificador Espaço-Temporal em Pseudo-Imagens (CETPI) projeta o histórico do esqueleto sob a forma de uma matriz densa espaço-temporal.

O processo fundamenta-se no achatamento (*flattening*) das dimensões estritamente espaciais e de atributos de cada nó do esqueleto, convertendo a geometria multiaxial de um frame em um vetor de características linearizado. Em seguida, esses vetores gerados a cada instante são empilhados cronologicamente ao longo do eixo vertical. O resultado final consiste em uma pseudo-imagem bidimensional estruturada com dimensões $(Frames, Características\ Achatadas)$, conforme apresentado na Figura 3. Nesse formato, o tempo atua como o eixo das ordenadas (altura) e as articulações mapeadas dispõem-se no eixo das abscissas (largura), permitindo que os filtros de convolução 2D realizem varreduras simultâneas sobre a textura e a assinatura espaço-temporal do movimento.

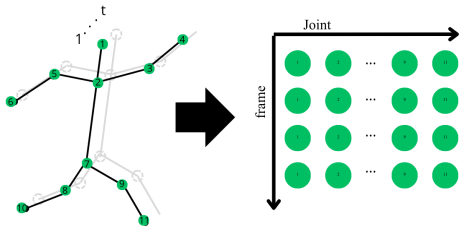


Figura 3. Exemplo de representação pseudo-imagem para um esqueleto 3D. O tempo é acumulado verticalmente e os atributos espaciais são achatados horizontalmente, compondo uma matriz convolucional 2D.

3.2. FitTube

Para a validação experimental da ferramenta, construiu-se o conjunto de dados (*dataset*) denominado *FitTube*, composto por trechos de vídeo de exercícios físicos extraídos do YouTube. O processo de amostragem e filtragem seguiu um protocolo estrito baseado em critérios predefinidos de elegibilidade, resultando em um *corpus* final de $N = 357$ vídeos.

3.2.1. Critérios de Inclusão e Seleção

A seleção dos trechos de vídeo obedeceu aos seguintes requisitos técnicos e de qualidade visual:

1. **Resolução espacial mínima:** 720p (1280×720 pixels);
2. **Taxa de quadros:** ≥ 24 quadros por segundo (*frames per second* – fps);
3. **Enquadramento corporal:** visibilidade completa da estrutura corporal do praticante durante a execução do exercício;
4. **Integridade visual:** ausência de oclusões significativas ou interrupções severas no campo de visão.

3.2.2. Taxonomia dos Exercícios

O *dataset* engloba 30 modalidades distintas de exercícios físicos, agrupadas em três classes taxonômicas com base no grupo muscular primário solicitado:

- **Membros Superiores:** flexão convencional (*push-up*), flexão diamante, flexão inclinada, flexão declinada, flexão pliométrica, flexão hindu, flexão com rotação, flexão *pseudo planche* e *dips* em banco.
- **Membros Inferiores:** agachamento isométrico (*wall sit*), agachamento pistola (*pistol squat*), agachamento búlgaro (*bulgarian split squat*), *lunge* alternado, *lunge* com salto, *lunge* reverso, *calf raise* unilateral, agachamento com deslocamento e *lunge* isométrico.
- **Abdominais (*Core*):** prancha frontal, prancha lateral, prancha com toque no ombro, prancha com rotação, *mountain climbers*, elevação de pernas, elevação de joelhos, abdominais bicicleta, torção russa, abdominais curtos, abdominais *jackknife* e prancha alta com elevação de braço.

3.2.3. Esquema de Anotação e Extração de Atributos

Cada segmento de vídeo foi submetido a um esquema de anotação estruturado composto pelos seguintes metadados:

1. **Identificador único:** código alfanumérico exclusivo para rastreabilidade individual do segmento;
2. **Coordenadas temporais:** marcações de tempo (*timestamps*) referentes ao início e término da execução do exercício;
3. **Classificação cinemática binária:** categorização da estabilidade do enquadramento da câmera durante o registro:
    - **Estático:** ausência de variação angular ou deslocamento da câmera durante a gravação;
    - **Dinâmico:** presença de movimentação, rotação ou alteração do ângulo de filmagem.
4. **Classe taxonômica:** rótulo anatômico-motor do exercício (*membros superiores*, *membros inferiores* ou *abdominais*).

Após a etapa de amostragem e rotulagem dos vídeos, procedeu-se à extração automática das coordenadas espaciais dos pontos de articulação corporal (*keypoints* do esqueleto), quadro a quadro, utilizando a biblioteca *MediaPipe Pose*. Os dados finais obtidos e a caracterização detalhada do conjunto de dados serão apresentados na Seção de Resultados e Discussão4.

4. Resultados e Discussão

Nesta seção, discutem-se os resultados quantitativos obtidos na avaliação do *framework 4SkeletonFeat* sob a ótica do estado da arte em reconhecimento de ações humanas (HAR) e biomecânica. O *benchmark*, realizado com 357 vídeos do *dataset* FitTube, revela o equilíbrio entre eficiência computacional e densidade informacional.

Os parâmetros de configuração adotados fixaram o tamanho do vocabulário em $K = 256$ centroides, utilizando o algoritmo K-Means com um orçamento estratificado de 60.000 instâncias de poses para o treinamento do dicionário. As estruturas de grafos foram exportadas no formato de coordenadas esparsas COO (`edge_index`). As métricas agregadas estão sintetizadas na Tabela 1.

Tabela 1. Métricas comparativas de desempenho computacional e armazenamento por paradigma no *framework 4SkeletonFeat* ($N = 357$ vídeos).

Paradigma Extraído	Tempo de Processamento (s)	Arquivos Gerados	Tamanho Total (MB)	Média por Arquivo (KB)
1. Joints Brutos (MediaPipe)	5227,74	357	57,37	164,56
2. Bone Vectors	47,01	357	21,21	60,84
3. Joint Motion(DCA)	7,33	357	54,01	154,92
4. Bone Motion (DCO)	2,76	357	19,70	56,51
5. Pose Tokens (K-Means)	79,81	357	0,09	0,26
6. Pseudo-Images (CETPI)	5,67	0[2]	0,00	0,00
7. Graph Structures (CTG)	5,88	357	57,61	165,25
Total do Pipeline	**5376,23 s (89,60 min)**	**2142**	**209,99 MB**	**–**

A análise do tempo de execução confirmou que a Extração de Joints Brutos consome 97% do tempo total do pipeline. Este gargalo é amplamente documentado na literatura, onde modelos como MediaPipe e OpenPose enfrentam latências significativas devido a pipelines multiestágio, geração de heatmaps e extração de características em alta resolução [Neupane et al. 2024]. Enquanto sistemas bottom-up (como OpenPose) sofrem com o custo de agrupamento de afinidade de partes, sistemas top-down (como os integrados ao framework) têm seu custo escalado pelo número de pessoas detectadas [Lan et al. 2024].

Destaca-se a velocidade de processamento, DCA, $(7,33$ s) e do DCO $(2,76$ s), demonstrando que a formulação matemática por diferenças finitas discretas implementada via operações vetorizadas nativas do NumPy minimiza o overhead computacional, esta eficiência é crucial, pois a literatura atual aponta que a representação multimodal — combinando juntas, ossos e movimento — é superior a abordagens unimodais [Chen et al. 2025, Cui et al. 2025, Shaikh et al. 2024]. A geração dos tokens demandou $79,81$ segundos, um tempo altamente competitivo considerando que este valor engloba tanto o treinamento do K-Means sobre 60.000 posturas, quanto a quantização e atribuição de IDs subsequentes para os 357 arquivos. Os valores obtidos para Pseudo-Images $(5,67$ s) e grafos $(5,88$ s) confirmaram a viabilidade de geração de tensores complexos em tempo de execução sub-segundo por clipe de movimento, o que corrobora estudos que indicam que métodos de processamento sobre poses extraídas podem atingir de 117 a mais de 9000 FPS [Rode et al. 2025].

Do ponto de vista de infraestrutura e armazenamento para Aprendizado Profundo, os resultados demonstram disparidades críticas entre as abordagens representacionais. As representações contínuas completas, como *Joints Brutos* $(57,37$ MB) e *Graph Structures* $(57,61$ MB), apresentaram o maior volume de armazenamento em disco. No caso dos grafos, o tamanho ligeiramente superior justifica-se pela necessidade de armazenar de forma persistente, não apenas os atributos quadridimensionais dos nós, mas também a matriz esparsa de conectividade `edge_index`. Os vetores anatômicos de ossos (*Bone Vectors* e *Bone Motion*) demonstraram uma redução natural de tamanho para ≈ 20 MB, decorrente do mapeamento otimizado de adjacências, que reduz a redundância espacial de nós periféricos. O achado mais impactante correlaciona-se ao paradigma de geração

[2]Matrizes estruturadas diretamente em cache/renderização de visualização.

de tokens. Ao converter a matriz contínua e multidimensional de geometrias articuladas em uma sequência unidimensional de índices discretos, o tamanho total de memória em disco despencou para $0,09$ MB para os 357 arquivos, registrando uma média de $0,26$ KB por arquivo.

Essa compactação representa uma economia de armazenamento de $99,84\%$ em relação à pose bruta original. Esse fenômeno valida cientificamente a eficiência da tokenização quantizada para cenários de Big Data e treinamento de arquiteturas generativas ou preditivas baseadas em Transformers autorregressivos, eliminando restrições de largura de banda de memória (*memory bandwidth*) e custos de armazenamento em larga escala.

A capacidade do framework proposto, *4SkeletonFeat*, de gerar Estruturas de Grafos (COO) e Pseudo-Imagens alinha-se às lideranças arquiteturais atuais. Enquanto as GCNs (Graph Convolutional Networks) tornaram-se o padrão por tratarem o corpo como um grafo natural de juntas e ossos, as pseudo-imagens permitem o reaproveitamento de backbones CNN 2D altamente otimizados [Hu et al. 2024, Cui et al. 2025].

A literatura sugere que os modelos mais fortes de 2024-2025 são frequentemente híbridos, combinando fluxos de grafos para topologia local com fluxos de Transformers para contexto global [Chen et al. 2025, Pan et al. 2024]. Ao fornecer todas essas representações a partir de uma única fonte de dados, o framework permite que pesquisadores explorem a fusão de modalidades, que comprovadamente supera sistemas de fluxo único ao compensar a falta de contexto ambiental dos esqueletos com a riqueza dinâmica do movimento.

5. Conclusão

Este trabalho apresentou o *4SkeletonFeat*, um *framework* modular projetado para integrar algoritmos de estimativa de pose a múltiplos paradigmas de modelagem para Aprendizado de Máquina e Aprendizado Profundo. A arquitetura desenvolvida traduz movimentos humanos em estruturas otimizadas para redes recorrentes (sequências), redes convolucionais (pseudo-imagens), redes em grafos (matrizes COO) e modelos de atenção (tokens discretos).

Os resultados experimentais validaram a eficiência da abordagem. Embora a extração inicial de articulações a partir do vídeo bruto constitua o principal gargalo computacional devido à inferência quadro a quadro, os módulos internos de engenharia biomecânica e estruturação de tensores executaram com elevada rapidez via vetorização matemática, consumindo menos de 3% do tempo total do *pipeline*. Ademais, a estratégia de quantização vetorial proporcionou alta taxa de compressão, reduzindo o espaço em disco em mais de $99,8\%$ e fornecendo uma representação simbólica leve para modelos fundamentais.

Em trabalhos futuros, pretende-se estender o *framework* para oferecer suporte a *pipelines* de processamento distribuído, integrando-o a fluxos de captura em tempo real (*edge computing*). Adicionalmente, planeja-se avaliar o impacto de cada um dos sete paradigmas gerados na acurácia final de modelos de reconhecimento de ações humanas em *benchmarks* públicos de larga escala, como o NTU RGB+D. Por fim, almeja-se estabelecer um *benchmark* próprio para o FitTube, avaliando diversos modelos de aprendizado de máquina.

Agradecimentos

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Programa de Excelência Acadêmica (PROEX), através da concessão de bolsa de estudo (Processo nº 88887.082808/2024-00).

Referências

Abdelkawy, A., Ali, A., and Farag, A. A. (2024). Epam-net: An efficient pose-driven attention-guided multimodal network for video action recognition. *Neurocomputing*, 633:129781.

Alomar, K., Aysel, H. I., and Cai, X. (2024). Cnns, rnns and transformers in human action recognition: a survey and a hybrid model. *Artificial Intelligence Review*, 58.

Chen, D., Chen, M., Wu, P., Wu, M., Zhang, T., and Li, C. (2025). Two-stream spatio-temporal gcn-transformer networks for skeleton-based action recognition. *Scientific Reports*, 15.

Cui, H., Huang, R., Zhang, R., and Hayama, T. (2025). Dstsa-gcn: Advancing skeleton-based gesture recognition with semantic-aware spatio-temporal topology modeling. *Neurocomputing*, 637:130066.

Gao, Z., Chen, J., Liu, Y., Jin, Y., and Tian, D. (2025). A systematic survey on human pose estimation: upstream and downstream tasks, approaches, lightweight models, and prospects. *Artificial Intelligence Review*, 58.

Halmich, C., Höschler, L., Schranz, C., and Borgelt, C. (2025). Data augmentation of time-series data in human movement biomechanics: A scoping review. *PLOS One*, 20.

Hu, H., Fang, Y., Han, M., and Qi, X. (2024). Multi-scale adaptive graph convolution network for skeleton-based action recognition. *IEEE Access*, 12:16868–16880.

Lan, G., Wu, Y., and Hao, Q. (2024). Dir-bhrnet: A lightweight network for real-time vision-based multiperson pose estimation on smartphones. *IEEE Transactions on Industrial Informatics*, 20:12533–12541.

Liu, M., Liu, H., Hu, Q., Ren, B., Yuan, J., Lin, J., and Wen, J. (2025). 3d skeleton-based action recognition: A review.

Neupane, R. B., Li, K., and Boka, T. F. (2024). A survey on deep 3d human pose estimation. *Artificial Intelligence Review*, 58.

Pan, L., Lu, J., and Tang, X. (2024). Spatial-temporal graph neural ode networks for skeleton-based action recognition. *Scientific Reports*, 14.

Pellano, K. N., Strümke, I., and Ihlen, E. A. F. (2024). From movements to metrics: Evaluating explainable ai methods in skeleton-based human activity recognition. *Sensors (Basel, Switzerland)*, 24.

Rode, D., Dunkel, A., Willi, R., Wolf, P., Xiloyannis, M., and Riener, R. (2025). Assessment of monocular human pose estimation models for clinical movement analysis. *Scientific Reports*, 15.

Shaikh, M. B., Chai, D., Islam, S. M. S., and Akhtar, N. (2024). From cnns to transformers in multimodal human action recognition: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20:1 – 24.

Tu, Z., Zhang, Z., Gong, J., Yuan, J., and Du, B. (2025). Informative sample selection model for skeleton-based action recognition with limited training samples. *IEEE Transactions on Image Processing*, 34:7335–7346.

Wang, H., Weng, W., Wang, J., Zhao, F., Xie, G.-S., Geng, X., and Wang, L. (2026). Foundation model for skeleton-based human action understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):47–61.

Xiang, L. and Wang, Z. (2024). Joint mixing data augmentation for skeleton-based action recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21:1 – 24.

Xie, D., Zhang, X., Gao, X., Zhao, H., and Du, D. (2025). Maf-net: A multimodal data fusion approach for human action recognition. *PLOS One*, 20.

Yan, J., Zhang, X., Tan, C., and Li, D. (2026). Skelformer: An adaptive hierarchical transformer-based approach on skeleton graphs for human action recognition in video sequences. *PLOS ONE*, 21(1):1–20.

Zhang, J., Lin, L., Yang, S., and Liu, J. (2024). Self-supervised skeleton-based action representation learning: A benchmark and beyond. *International Journal of Computer Vision*, 134.