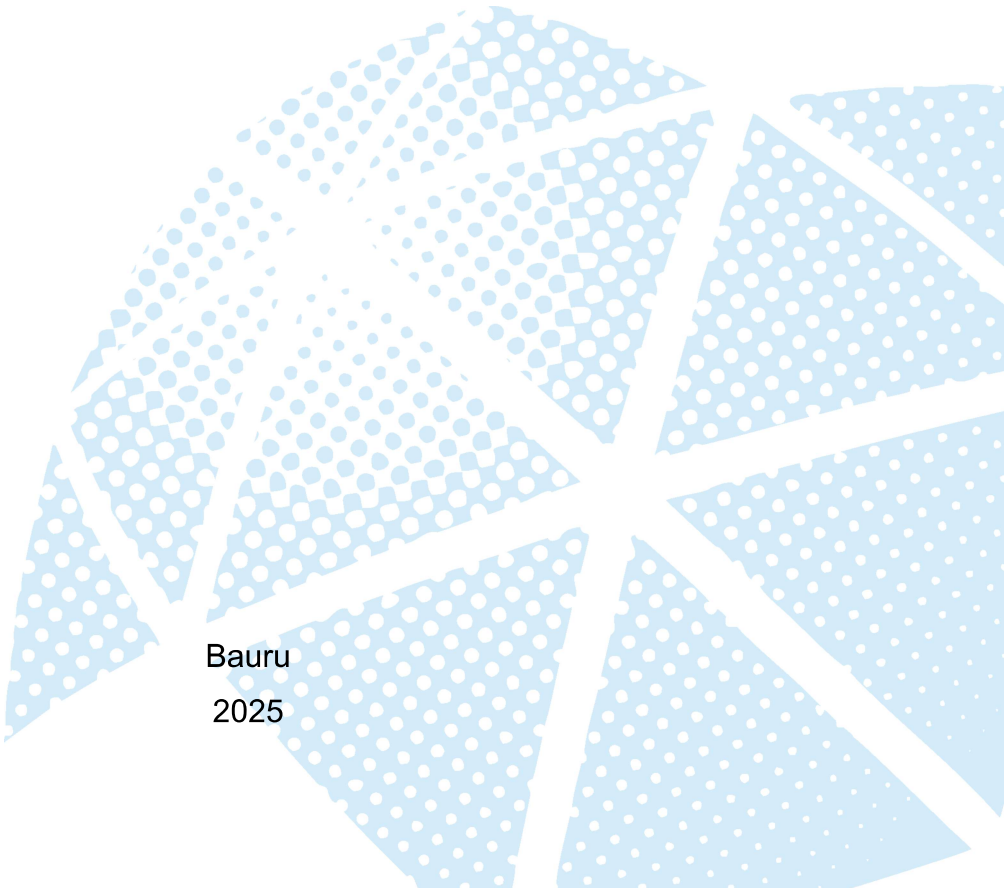


UNIVERSIDADE ESTADUAL PAULISTA - UNESP
Faculdade de Ciências - Campus de Bauru

Daniel Ricardo dos Santos Jangua

**Reconhecimento de Ações Humanas Utilizando Poses 3D Estimadas a Partir
de Poses 2D**

Bauru
2025



Daniel Ricardo dos Santos Jangua

**Reconhecimento de Ações Humanas Utilizando Poses 3D Estimadas a Partir
de Poses 2D**

Dissertação apresentada à Universidade Estadual Paulista (UNESP), Faculdade de Ciências, Bauru, para obtenção do título de Mestre em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador: Prof. Dr. Aparecido Nilceu Marana

Bauru
2025

Jangua, Daniel Ricardo dos Santos. .
Reconhecimento de Ações Humanas Utilizando Poses 3D Estimadas a Partir
de Poses 2D / Daniel Ricardo dos Santos Jangua. – Bauru, 2025
91 f. : il., tabs.

Orientador: Prof. Associado Aparecido Nilceu Marana
Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita
Filho", Faculdade de Ciências, Bauru, 2025.

1. Reconhecimento de Ações Humanas. 2. Estimção de Poses 3D. 3. Visão
Computacional. I. Marana, Aparecido Nilceu II. Universidade Estadual Paulista "Júlio
de Mesquita Filho", Faculdade de Ciências. III. Reconhecimento de Ações Humanas
Utilizando Poses 3D Estimadas a Partir de Poses 2D.

CDU – 518.72:76

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE DANIEL RICARDO DOS SANTOS JANGUA, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 21 dias do mês de agosto do ano de 2025, às 14h, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de DANIEL RICARDO DOS SANTOS JANGUA intitulada Reconhecimento de Ações Humanas Utilizando Poses 3D Estimadas a Partir de Poses 2D. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. APARECIDO NILCEU MARANA (Orientador - Participação Virtual) do Departamento de Computação, Faculdade de Ciências, Unesp, Bauru, SP, Prof. Dr. LUÍS ANTONIO DE SOUZA JÚNIOR (Participação Virtual) do Dep. de Informática, Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, ES, Profa. Dra ROBERTA SPOLON (Participação Virtual) do Departamento de Computação, Faculdade de Ciências, Unesp, Bauru, SP. Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma virtual, o discente recebeu o conceito final **Aprovado**. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo Presidente da Comissão Examinadora.



Documento assinado digitalmente

APARECIDO NILCEU MARANA

Data: 21/08/2025 17:53:59-0300

Verifique em <https://validar.iti.gov.br>

Prof. Dr. APARECIDO NILCEU MARANA

Dedico esta dissertação àqueles que caminharam ao meu lado durante essa jornada.

Agradecimentos

Agradeço à minha família, em especial aos meus pais, que me serviram de apoio e inspiração durante todo o mestrado.

Agradeço aos meus amigos que tornaram meu caminho mais fácil e cheio de alegria.

Agradeço aos meus professores, em especial ao meu orientador Prof. Associado Aparecido Nilceu Marana que me acompanhou desde o meu primeiro contato com a pesquisa científica, ainda nos primeiros anos da graduação.

Agradeço aos meus companheiros de laboratório, em especial ao João Renato Ribeiro Manesco, por sua valiosa colaboração.

Agradeço à UNESP e ao PPGCC-UNESP pela infraestrutura e recursos disponibilizados para a realização deste trabalho.

Mesmo as noites totalmente sem estrelas podem anunciar a aurora de uma grande realizaão.

Martin Luther King Jr.

Resumo

O reconhecimento de ações humanas baseado em poses tem sido uma área de pesquisa ativa nas áreas de Visão Computacional e Aprendizado de Máquina. Apesar dos avanços já alcançados, ainda existem diversos obstáculos a serem superados, principalmente em relação à estimação das posições do corpo humano em vídeo, que pode ser muito desafiadora em cenários onde ocorrem obstruções ou sobreposições por outros indivíduos ou objetos. Para estes casos, seria interessante estimar as coordenadas do corpo humano em três dimensões, de forma a obter uma representação mais precisa. No entanto, a implementação, a usabilidade e a manutenção de sensores que captam profundidade podem não ser viáveis em muitas aplicações, nas quais a captura dos dados brutos se baseia apenas em câmeras RGB simples. Uma alternativa seria estimar as poses no espaço 3D a partir da triangulação de poses 2D capturadas a partir de dois ou mais ângulos de visão de maneira simultânea. Porém, essa solução é desafiadora quando os parâmetros da câmera são desconhecidos ou torna-se inexequível quando apenas uma visão de câmera está disponível. Este trabalho tem como objetivo propor uma abordagem para reconhecimento de ações humanas com base em poses 3D estimadas a partir de uma única câmera RGB monocular. A abordagem proposta utiliza o DUA, um método desenvolvido recentemente para estimar poses 3D a partir de poses 2D usando adaptação de domínio. Para extração de características e classificação, dois métodos foram avaliados neste trabalho: 1DCNN e GCN. Experimentos realizados em um subconjunto de ações humanas do conjunto de dados NTU RGB+D mostraram que a abordagem proposta, baseada em poses 3D, foi exitosa pois obteve uma acurácia maior do que ao se utilizar poses 2D. No cenário de ações limitadas, os resultados alcançaram 79.39% (*X-Sub*) e 78.81% (*X-View*) com poses 2D, contra 80.97% (*X-Sub*) e 81.42% (*X-View*) com poses 3D. Já quando considerado o conjunto completo de ações, o desempenho foi de 59.56% (*X-Sub*) e 67.66% (*X-View*) em 2D, enquanto as poses 3D obtiveram 62.04% (*X-Sub*) e 66.68% (*X-View*). A análise comparativa com métodos do estado da arte indicou que, embora a abordagem proposta apresente desempenho competitivo em cenários reduzidos, sua eficácia diminui ao lidar com conjuntos de ações mais amplos devido à expressividade limitada da topologia esquelética utilizada. A inclusão de variantes baseadas em movimento demonstrou potencial para enriquecer a representação espacial, resultando no melhor desempenho no subconjunto de 10 ações, validando a integração entre estimação de pose 3D monocular e modelos de reconhecimento baseados em grafos como alternativa viável para cenários sem sensores dedicados.

Palavras-chave: Estimação de Poses 3D; Reconhecimento de Ações Humanas; Visão Computacional.

Abstract

Human action recognition based on poses has been an active research area in the fields of Computer Vision and Machine Learning. Despite the advances already achieved, there are still several obstacles to overcome, mainly regarding the estimation of human body positions in video, which can be very challenging in scenarios with occlusions or overlaps caused by other individuals or objects. In such cases, it would be interesting to estimate the human body coordinates in three dimensions, in order to obtain a more accurate representation. However, the implementation, usability, and maintenance of sensors that capture depth may not be feasible in many applications, where data acquisition relies solely on simple RGB cameras. An alternative would be to estimate 3D poses through the triangulation of 2D poses captured simultaneously from two or more viewpoints. Nevertheless, this solution is challenging when camera parameters are unknown, or becomes impractical when only a single camera view is available. This work aims to propose an approach for human action recognition based on 3D poses estimated from a single monocular RGB camera. The proposed approach employs DUA, a recently developed method to estimate 3D poses from 2D poses using domain adaptation. For feature extraction and classification, two methods were evaluated in this work: 1DCNN and GCN. Experiments conducted on a subset of human actions from the NTU RGB+D dataset showed that the proposed approach, based on 3D poses, was successful since it achieved higher accuracy compared to using 2D poses. In the limited action scenario, the results reached 79.39% (*X-Sub*) and 78.81% (*X-View*) with 2D poses, against 80.97% (*X-Sub*) and 81.42% (*X-View*) with 3D poses. When considering the full set of actions, the performance was 59.56% (*X-Sub*) and 67.66% (*X-View*) for 2D poses, while 3D poses achieved 62.04% (*X-Sub*) and 66.68% (*X-View*). The comparative analysis with state-of-the-art methods indicated that, although the proposed approach presents competitive performance in reduced scenarios, its effectiveness decreases when dealing with broader sets of actions due to the limited expressiveness of the skeletal topology employed. The inclusion of motion-based variants demonstrated potential to enrich the spatial representation, resulting in the best performance in the 10-action subset, validating the integration between monocular 3D pose estimation and graph-based recognition models as a viable alternative for scenarios without dedicated sensors.

Keywords: 3D Pose Estimation; Human Action Recognition; Computer Vision.

Lista de ilustrações

Figura 1	– Exemplos de aplicações do mundo real utilizando técnicas de reconhecimento de ações humanas. Da esquerda para a direita: (i) interação entre humano e robô, (ii) jogos interativos voltados para o entretenimento e (iii) carros com navegação autônoma.	18
Figura 2	– Exemplo de aplicação do OpenPose, um método <i>bottom-up</i> de estimação de poses 2D com múltiplas pessoas.	20
Figura 3	– Estimação de poses 3D a partir de imagens monoculares.	21
Figura 4	– Triangulação de um ponto 3D a partir de duas observações 2D.	22
Figura 5	– Modelos de poses 2D usados pelos métodos <i>OpenPose</i> (a) e <i>PifPaf</i> (b).	28
Figura 6	– <i>Pipeline</i> do método OpenPose: (a) entrada da imagem completa na CNN; (b) geração de mapas de confiança para detecção das partes do corpo; (c) geração de PAFs para associação das partes; (d) pareamento bipartido para associar candidatos a partes; (e) montagem das poses completas de todos os indivíduos	29
Figura 7	– Arquitetura da CNN multiestágio: (a) primeiros estágios preveem PAFs L_t ; (b) últimos estágios preveem mapas de confiança S_t ; (c) previsões e características da imagem de cada estágio são concatenadas para o estágio seguinte; (d) convoluções originais de <i>kernel</i> 7 são substituídas por três camadas de <i>kernel</i> 3 concatenadas.	29
Figura 8	– Arquitetura do modelo: a entrada é uma imagem ($H \times W$) com três canais de cor ($\times 3$). O <i>encoder</i> da rede neural gera os campos PIF e PAF com 17 e 19 canais, respectivamente, e operações com <i>stride</i> 2 são indicadas por $//2$. O <i>decoder</i> converte os campos PIF e PAF em estimativas de pose com 17 articulações, cada uma representada por coordenadas x e y e um score de confiança.	30
Figura 9	– Exemplos de resultados obtidos pelo método HMR.	32
Figura 10	– Comparação qualitativa dos resultados obtidos pelo VIBE (linha superior) e HMR (KANAZAWA et al., 2018) (linha inferior).	33
Figura 11	– Estimação da pose 3D a partir de diferentes ângulos de visão 2D.	34
Figura 12	– Fluxo do método ST-GCN: estimação de poses em vídeo, construção do grafo espaço-temporal e classificação das ações via <i>Softmax</i> (YAN et al., 2018).	45
Figura 13	– Estratégias de particionamento propostas para a construção das operações de convolução: (a) quadro de entrada com juntas corporais (azul) e campo receptivo (vermelho tracejado); (b) particionamento por rótulo único; (c) particionamento por distância; (d) particionamento por configuração espacial.	47

Figura 14 – Visão geral do funcionamento do método de ST-GCN estendido (WANG et al., 2022a).	49
Figura 15 – Grafo espaço-temporal de uma sequência de poses proposto por Yan et al. (2018). As conexões físicas são representadas em azul, e as temporais em verde.	50
Figura 16 – Exemplo de um grafo estendido com conexões naturais em preto e conexões entre nós não adjacentes em verde (à esquerda), e a composição do grafo espaço-temporal com as conexões temporais em laranja (à direita) (WANG et al., 2022a).	50
Figura 17 – Esquema de particionamento estendido: nó τ em vermelho, nós ξ em verde, nós ϕ em laranja, nós λ em roxo e nós μ em amarelo (WANG et al., 2022a).	51
Figura 18 – Comparação entre os grafos espaço-temporais utilizado pelo método ST-GCN (esquerda) (YAN et al., 2018) e <i>Hybridnet</i> (direita) (YANG et al., 2023).	53
Figura 19 – (Esquerda) Atenção por Produto Escalar Escalonado. (Direita) Atenção Multi-Cabeça composta por várias camadas de atenção em paralelo.	56
Figura 20 – A camada de atenção gera vetores de contexto para a modelagem sequência a sequência. Na autoatenção, múltiplas cabeças são criadas com diferentes inicializações e suas saídas são concatenadas na atenção multicabeça.	57
Figura 21 – Visão geral do modelo de Dosovitskiy et al. (2020): a imagem é dividida em <i>patches</i> de tamanho fixo, linearmente embutidos com <i>embeddings</i> posicionais e processados por um <i>Transformer encoder</i> . Para classificação, utiliza-se um <i>token</i> adicional aprendível.	58
Figura 22 – Visão geral da arquitetura proposta por Mazzia et al. (2022): estimativas de pose são projetadas linearmente, combinadas ao <i>class token</i> e processadas pelo <i>Transformer encoder</i> com <i>embeddings</i> posicionais aprendíveis. Apenas o <i>class token</i> de saída passa por uma MLP para gerar a predição final da ação.	60
Figura 23 – Arquitetura do RGBSformer: dois fluxos complementares, RGB (lento) para características espaciais e esqueleto (rápido) para informação temporal, fundidos em <i>Transformer</i> para a predição de ações (SHI et al., 2023).	63
Figura 24 – Diagrama em blocos da abordagem proposta nesta dissertação para reconhecimento de ações humanas com poses 3D estimadas a partir de imagens RGB monoculares.	65
Figura 25 – Diagrama do método proposto para reconhecimento de ações baseado em ângulos e distâncias das juntas e usando um classificador 1D CNN.	67
Figura 26 – Arquitetura utilizada para o modelo de ST-GCN Motion Based (ST-GCNMB).	73

Figura 27 – <i>Frames</i> extraídos do conjunto de dados NTU RGB+D. As primeiras quatro linhas mostram a variedade de indivíduos e visualizações de câmeras. A quinta linha representa a variação intraclasse das ações. A última linha ilustra as modalidades RGB, RGB+juntas, profundidade, profundidade+juntas e IR de um <i>frame</i> de amostra.	75
Figura 28 – <i>Frames</i> extraídos do conjunto de dados NTU RGB+D que ilustram as 10 ações avaliadas neste trabalho. Da esquerda para a direita, de cima para baixo: (a) pegar algo do chão; (b) arremessar; (c) bater palmas; (d) acenar; (e) chutar; (f) colocar a mão no bolso; (g) pular com um pé; (h) pular para cima; (i) esfregar as duas mãos; (j) fazer o sinal de pare.	77
Figura 29 – Matrizes de confusão geradas pelos modelos ST-GCN utilizando poses 2D.	80
Figura 30 – Matrizes de confusão geradas pelos modelos ST-GCN utilizando poses 3D.	81

Lista de tabelas

Tabela 1 – Acurácia Top-1 no conjunto de dados NTU RGB+D.	39
Tabela 2 – Resultados obtidos pelo modelo de Yan et al. (2018).	48
Tabela 3 – Resultados obtidos pelo modelo de Wang et al. (2022a).	51
Tabela 4 – Resultados obtidos pelos modelos <i>Hybridnet</i> de Yang et al. (2023).	53
Tabela 5 – Configurações do modelo <i>Action Transformer</i>	62
Tabela 6 – Resultados obtidos pelo modelo <i>Action Transformer</i> (OpenPose).	62
Tabela 7 – Resultados obtidos pelo modelo de Shi et al. (2023).	64
Tabela 8 – Acurácia dos métodos avaliados utilizando o conjunto de 10 ações.	78
Tabela 9 – Mapeamento do conjunto de 10 ações.	79
Tabela 10 – Acurácias obtidas por métodos estado-da-arte e pelo método ST-GCN proposto neste trabalho, no conjunto de 40 ações cotidianas do conjunto de dados NTU RGB+D.	82

Lista de abreviaturas e siglas

1-NN	<i>1-Nearest Neighbour</i>
2D	2 Dimensões
2S-EGCN	<i>2-Stream Extended Graph Convolutional Network</i>
3D	3 Dimensões
DANN	<i>Domain-Adversarial Neural Networks</i>
DUA	<i>Domain Unified Approach</i>
CLS	<i>Token de Classe</i>
CNN	<i>Convolutional Neural Network</i>
CS	<i>Cross-subject</i>
CV	<i>Cross-view</i>
GAT	<i>Graph Attention Networks</i>
GCN	<i>Graph Convolutional Network</i>
HMR	<i>Human Mesh Recovery</i>
HOG	<i>Histogram of Oriented Gradients</i>
IR	Infravermelho
LSTM	<i>Long Short-Term Memory</i>
MLP	<i>Multi-layer Perceptron</i>
NPD	Distância de Pose Normalizada
PAF	<i>Part Association Fields</i>
PIF	<i>Part Intensity Fields</i>
PLN	Processamento de Linguagem Natural
RGB	<i>Red, Green and Blue</i>
RNN	<i>Recurrent Neural Networks</i>

SfM	<i>Structure-from-Motion</i>
SMPL	<i>Skinned Multi-Person Linear Model</i>
ST-GCN	<i>Spatio-Temporal Graph Convolutional Network</i>
ST-GCNMB	<i>Spatio-Temporal Graph Convolutional Network Motion-Based</i>
TSC	<i>Temporal Series Classification</i>
ViT	<i>Vision Transformer</i>
VIBE	<i>Video Inference for Human Body Pose and Shape Estimation</i>
X-Sub	<i>Cross-subject</i>
X-View	<i>Cross-view</i>

Sumário

1	INTRODUÇÃO	18
1.1	Hipótese	23
1.2	Objetivo	23
1.3	Contribuição Esperada	23
1.4	Organização da Dissertação	23
2	ESTIMAÇÃO DE POSE HUMANA	25
2.1	Introdução	25
2.2	Métodos de Estimação de Poses 2D	27
2.2.1	<i>OpenPose</i>	28
2.2.2	<i>PifPaf</i>	29
2.3	Métodos de Estimação de Poses 3D	30
2.3.1	<i>Human Mesh Recovery</i>	31
2.3.2	<i>Video Inference for Human Body Pose and Shape Estimation</i>	32
2.3.3	Aprendizado de Poses 3D a partir de Múltiplas Visões	33
2.3.4	Aprendizado de Poses 3D a partir de Imagens Monoculares	35
2.4	Método Híbrido para Estimação de Poses 3D e Reconhecimento de Ações	37
2.5	Transformer-GCN com Aprendizado de Representações Contextualizadas	38
3	RECONHECIMENTO DE AÇÕES HUMANAS	40
3.1	Introdução	40
3.2	Aplicações no Mundo Real	41
3.3	Desafios	42
3.3.1	Variabilidade Inter e Intra-classe	42
3.3.2	Capacidade dos Sensores	43
3.3.3	Condições Ambientais	43
3.3.4	Acesso à Conjuntos de Dados	44
3.3.5	Limitações Computacionais	44
3.4	Métodos para Reconhecimento de Ações Humanas	44
3.4.1	Modelo Baseado em Grafo Espaço-Temporal	45
3.4.2	Extensão do Modelo Baseado em Grafo Espaço-Temporal	48
3.4.3	Modelo de Arquitetura Híbrida	52
3.4.4	Abordagens Baseadas em <i>Transformers</i>	54
3.4.5	<i>Transformers</i> na Visão Computacional	56

3.4.6	<i>Transformers</i> para Reconhecimento de Ações Humanas	59
4	ABORDAGEM PROPOSTA	65
4.1	Visão Geral da Abordagem	65
4.2	Extração de Características	66
4.2.1	Descritores Baseados em Ângulos e Distâncias das Juntas	66
4.2.2	Descritores Baseados em Ângulos Medidos em Quatérnios	68
4.2.3	Descritores Baseados em Grafos	68
4.3	Reconhecimento da Ação	69
4.3.1	1D CNN	69
4.3.2	Rede de Convolução Espaço-Temporal em Grafos	70
4.3.3	Rede de Convolução Espaço-Temporal em Grafos com Princípios de Atenção	71
5	RESULTADOS E DISCUSSÃO	74
5.1	Base de Dados	74
5.2	Cenários dos Experimentos	76
5.3	Métrica de Avaliação	76
5.4	Resultados Obtidos	78
6	CONCLUSÃO	83
	REFERÊNCIAS	85

1 Introdução

Em todos os setores da vida cotidiana, não importa o quão trivial seja o objetivo, uma ação individual, ou uma sequência de ações, deve ser realizada para cumprir com um determinado propósito. Seja para executar alguma tarefa no trabalho, ou até mesmo completar uma atividade física, um indivíduo responde ao ambiente usando seu torso e membros. Por meio da visão humana, é possível compreender facilmente a ação e o propósito de um ator apenas observando seus movimentos corporais e as interações com objetos da cena, no entanto, essa não é uma tarefa tão simples para uma inteligência artificial (KONG; FU, 2022).

O reconhecimento de ações é um dos principais desafios na área da visão computacional, e se baseia em atribuir a uma máquina a capacidade de reconhecer e compreender ações humanas (SUN et al., 2023). Este é um campo de estudo de grande importância para diversas aplicações no mundo real, em cenários onde se deseja extinguir ou diminuir a necessidade de uma supervisão humana, como em sistemas de vigilância por câmeras e navegação autônoma, além de diversas outras aplicações voltadas para o entretenimento e o aprimoramento da interação humano-máquina.

A Figura 1 ilustra alguns exemplos de aplicações do reconhecimento de ações humanas. Na imagem da esquerda é possível observar uma interação entre um humano e um robô, na imagem do centro, uma pessoa controla ações dentro de uma aplicação de computador por meio de seus gestos, já na imagem da direita, um carro autônomo aguarda a passagem de um ciclista e um pedestre para que possa seguir até seu destino. Em todos esses casos, é essencial que a máquina seja capaz de reconhecer ações humanas para atuar em resposta à elas da maneira mais adequada.

Figura 1 – Exemplos de aplicações do mundo real utilizando técnicas de reconhecimento de ações humanas. Da esquerda para a direita: (i) interação entre humano e robô, (ii) jogos interativos voltados para o entretenimento e (iii) carros com navegação autônoma.



Fonte: Kong and Fu (2022).

Apesar de não ser um campo de estudo novo da visão computacional quando comparado

à geração de imagens artificiais, por exemplo, o reconhecimento de ações humanas ainda se mostra em evidência, apresentando diversas possibilidades de evolução, principalmente nos últimos anos, com a disponibilidade de maior poder computacional e avanços nos métodos de processamento de imagens (KONG; FU, 2022). Esse constante interesse no assunto por parte da comunidade científica se dá uma vez que o aprimoramento da capacidade das máquinas de reconhecerem e compreenderem ações humanas é um fator crítico para o futuro da aplicação de sistemas de inteligência artificial no cotidiano.

De acordo com Sun et al. (2023), os primeiros trabalhos pautados no reconhecimento de ações humanas utilizavam como entrada de dados vídeos RGB (vermelho, verde e azul, do inglês, *Red, Green and Blue*) ou em escala de cinza, devido à popularidade e fácil acesso desse tipo de dado na época. No entanto, com o desenvolvimento de diferentes sensores mais baratos e precisos, nos últimos anos têm sido produzidos trabalhos baseados nas mais diversas modalidades de dados, como esqueleto (ou poses), imagens com profundidade, nuvem de pontos, aceleração, radar e até mesmo *WiFi*.

Dentre as modalidades supracitadas, o reconhecimento de ações humanas baseado em poses vem se tornando notável nos últimos anos, pois apresenta uma solução compacta e naturalmente focada em ações (DUAN et al., 2022), uma vez que codifica de forma eficiente e compacta a trajetória das juntas do corpo humano durante a execução dos movimentos (SUN et al., 2023). De modo geral, nesta modalidade os esqueletos humanos são representados por meio de uma sequência de coordenadas para cada uma das juntas do corpo humano mapeadas no modelo do esqueleto, extraídas por métodos de estimação de pose. Uma vez que os esqueletos capturam apenas a informação relacionada à ação, esta modalidade apresenta soluções imunes à mudanças de contexto como variações de fundo do cenário e iluminação (DUAN et al., 2022). Por outro lado, não são ideais para o reconhecimento de ações nas quais o contexto da cena é determinante ou que envolvem interação com objetos, além de trabalhar com dados possivelmente ruidosos derivados de imprecisões no processo de estimação de pose (SUN et al., 2023), já que os métodos são sensíveis à obstruções e sobreposições do corpo na imagem, além de apresentar falsos positivos.

A estimação de pose é uma tarefa desafiadora de visão computacional que busca, a partir de imagens ou vídeos, estimar informações a respeito da pose humana, como a posição de juntas do corpo ou parâmetros da forma do corpo humano (LIU et al., 2022). Uma vez que não requer um conjunto complexo com múltiplas câmeras ou roupas especiais com marcadores, a estimação de pose se tornou essencial em diversas aplicações do mundo real. De acordo com sua dimensionalidade, os estimadores de pose podem ser classificados em 2D (duas dimensões) ou 3D (3 dimensões), neste caso adicionando informações de profundidade às coordenadas das juntas do esqueleto humano.

Além disso, os métodos de estimação de pose também podem ser específicos para imagens com uma única pessoa ou com múltiplas pessoas, que por sua vez ainda podem ser

divididos em duas categorias de acordo com a abordagem tomada para estimação das múltiplas poses na cena, sendo elas: (i) abordagem *top-down* (do topo para baixo), onde cada indivíduo na cena é tratado separadamente e o método de estimação de pose é aplicado para cada um dos indivíduos detectados; (ii) abordagem *bottom-up* (da base para cima), onde toda a imagem é considerada para a estimação das poses, inicialmente todas as juntas de todos os indivíduos são detectadas e em seguida são combinadas formando as poses individuais (LIU et al., 2022). Os estimadores de poses mais recentes e performáticos usualmente adotam a abordagem *bottom-up*. A Figura 2 mostra um exemplo de aplicação de um método *bottom-up* de estimação de poses 2D com múltiplas pessoas denominado OpenPose (CAO et al., 2018).

Figura 2 – Exemplo de aplicação do OpenPose, um método *bottom-up* de estimação de poses 2D com múltiplas pessoas.



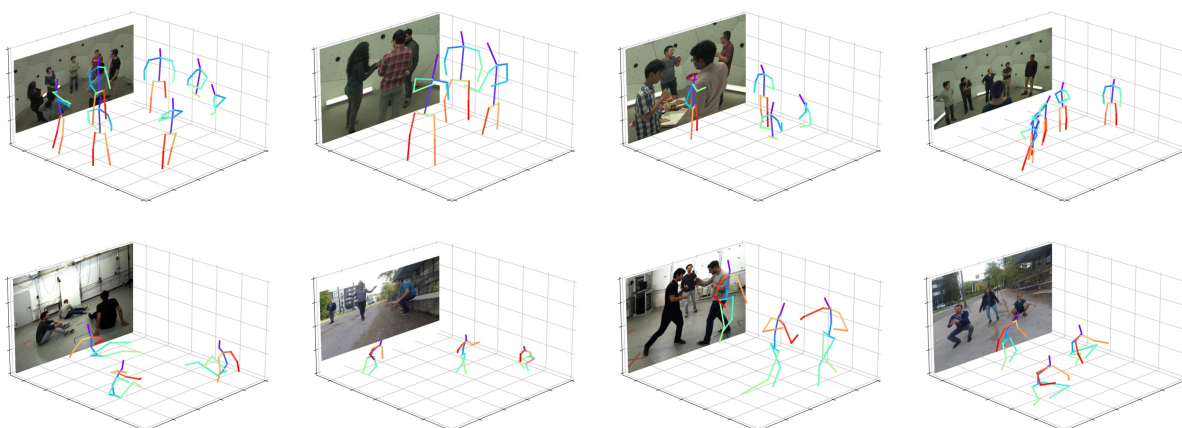
Fonte: Cao et al. (2018).

Os métodos de estimação de pose 3D podem ser divididos em dois grupos, de acordo com a abordagem tomada para obter a informação 3D: (i) métodos de ponta-a-ponta, que buscam obter as poses tridimensionais diretamente; (ii) métodos de dois estágios, que se baseiam em estimar as poses 3D a partir de poses 2D obtidas em um primeiro momento (WANG et al., 2022c).

Os métodos baseados em um único estágio foram estudados como um caminho inicial, no entanto os métodos de duas etapas se provaram mais vantajosos. Isso porque, ao se aproveitar da robustez de detectores de pose 2D já consolidados (como o OpenPose (CAO et al., 2018)), tais métodos permitem primeiro uma detecção precisa e econômica de poses 2D, que são abundantemente disponíveis e exigem menos dados que poses 3D. Em seguida, essas poses 2D são elevadas para o espaço 3D, beneficiando-se de recursos de adaptação que mantêm a precisão mesmo em cenários variados. Porém, essa abordagem depende criticamente da qualidade das poses 2D: em situações com oclusões severas, os pontos chave 2D imprecisos comprometem seriamente o resultado final em 3D, o que se confirma como um dos principais desafios da estratégia de duas etapas (KAID; BAÏNA, 2023).

Além disso, a estimação de poses 3D a partir de poses 2D apresenta outros desafios a serem superados, uma vez que busca reconstruir as localizações de juntas do corpo humano no espaço tri-dimensional a partir apenas de observações monoculares ou multioculares, que possuem um alto grau de ambiguidade em relação à profundidade nas imagens. Apesar dos desafios, esta é uma área de pesquisa de grande importância, principalmente para o reconhecimento de ações humanas, uma vez que consegue obter informações significativas a respeito da geometria do movimento do corpo humano, possibilitando uma análise mais detalhada das poses em relação aos métodos bi-dimensionais (ZHANG et al., 2022). A Figura 3 ilustra resultados obtidos pela estimação de poses 3D a partir de imagens monoculares (WANG et al., 2022c).

Figura 3 – Estimação de poses 3D a partir de imagens monoculares.

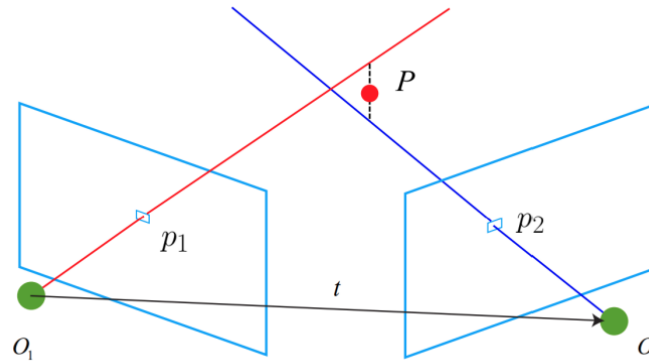


Fonte: Wang et al. (2022c).

Em contextos em que existam disponíveis dois ou mais ângulos de visão de um indivíduo executando determinada ação, seria possível utilizar métodos já consolidados para extração das poses 2D, e com essas informações, construir a pose 3D, estimando os pontos através de um processo de triangulação (HARTLEY; ZISSERMAN, 2004). A Figura 4 ilustra o processo de triangulação, onde o ponto P é obtido através de duas observações p_1 e p_2 , pelos observadores O_1 e O_2 , respectivamente.

Por outro lado, há situações em que apenas uma câmera RGB está disponível, resultando em imagens monoculares. Nesse contexto, ganham relevância os métodos que estimam poses humanas em 3D a partir de uma única imagem. Com o desenvolvimento das técnicas de estimação de poses 2D, que apresentam resultados consistentes por meio de modelos como o OpenPose (Cao et al. (2018)) e o PifPaf (Kreiss et al. (2019)), surgiram abordagens que buscam utilizar essas estimativas como base para a reconstrução tridimensional. Essas abordagens seguem um processo de dois estágios: primeiro, realiza-se a detecção das poses 2D; em seguida, essas poses são utilizadas como entrada para a recuperação da pose em 3D. A principal vantagem dessa estratégia está na disponibilidade de dados anotados em 2D, que permite o treinamento de modelos em diferentes cenários e viabiliza a estimação de poses em 3D sem

Figura 4 – Triangulação de um ponto 3D a partir de duas observações 2D.



Fonte: Hartley and Zisserman (2004).

a necessidade de sensores adicionais. Apesar disso, o problema apresenta limitações, já que a projeção bidimensional elimina as informações de profundidade e diferentes poses em 3D podem gerar projeções semelhantes em 2D. Ainda assim, trabalhos como o de Moreno-Noguer (2017), que utiliza matrizes de distâncias euclidianas entre articulações, e o de Manesco et al. (2023), que propõe uma abordagem unificada para estimação de poses 3D em cenários com diferentes domínios, mostram que é possível estimar poses em 3D com resultados satisfatórios mesmo com baixo custo computacional.

Uma vez obtidas as poses 3D, no contexto do reconhecimento de ações humanas baseado em poses, é necessário aplicar o método adequado, seja em questões de acurácia (eficácia) ou desempenho computacional (eficiência). Atualmente, existem diversos métodos estado-da-arte para reconhecimento de ações que são baseados em técnicas complexas e modelagens rebuscadas, como redes neurais de grafo ou *transformers* (MAZZIA et al., 2022; WANG et al., 2022a; YANG et al., 2023). Apesar de apresentarem ótimas acurácias, também apresentam um alto custo computacional, que pode se tornar um limitador para muitos casos de uso. Nesses casos, seria ideal o uso de técnicas mais simples, porém sem que a acurácia seja muito prejudicada em relação aos métodos estado-da-arte.

Este trabalho busca aplicar e avaliar um método simples e computacionalmente eficiente para o reconhecimento de ações humanas em 3D, sem a necessidade de sensores especializados para estimativa de profundidade, como o *Kinect*. Para isso, propõe-se a integração de uma técnica de estimativa de pose 3D invariável ao domínio, denominada *Domain Unified Approach* (DUA) (MANESCO et al., 2023), com modelos de reconhecimento de ações que operam diretamente sobre os esqueletos 3D estimados. A abordagem permite as seguintes contribuições: (i) o desenvolvimento de um sistema *end-to-end* que conecta a estimativa de pose 3D invariável ao domínio a métodos de reconhecimento de ações baseados nessas poses estimadas; e (ii) uma análise do desempenho e das limitações do método proposto, incluindo a identificação de seus principais aspectos positivos e a discussão de restrições, como aquelas relacionadas à quantidade reduzida de articulações presente nos modelos de estimativa de pose.

1.1 Hipótese

A estimação de poses 3D a partir de poses 2D capturadas a partir de uma visão monocular em diferentes ângulos de visão pode oferecer uma solução computacionalmente eficiente e simplificada para o reconhecimento de ações humanas em vídeos. A hipótese é de que esta abordagem, embora mais simples comparada aos métodos estado-da-arte, é capaz de fornecer uma acurácia comparável em reconhecimento de ações, sendo benéfica em cenários onde os recursos computacionais são limitados ou uma implementação mais simples é necessária.

1.2 Objetivo

O objetivo geral deste trabalho foi propor e avaliar uma abordagem para reconhecimento de ações humanas em vídeos que utiliza a estimativa de poses 3D derivada de poses 2D obtidas em um único ângulo de visão.

Inicialmente, almejou-se adaptar o método proposto por Jangua and Marana (2020) para o reconhecimento de marcha humana, com base em poses 2D, à tarefa de reconhecimento de ações humanas. Após, buscou-se aprimorar o processo de reconhecimento de ações humanas utilizando-se poses 3D estimadas por meio do método estado-da-arte denominado DUA, proposto por Manesco et al. (2023), e utilizando-se redes neurais de grafos, uma vez que estas permitem modelar naturalmente as articulações do esqueleto humano e propiciam um menor custo computacional.

1.3 Contribuição Esperada

Espera-se que este trabalho contribua para a área de reconhecimento de ações humanas ao fornecer um método computacionalmente eficiente e de fácil implementação. Esta contribuição tem o potencial de beneficiar várias aplicações de visão computacional, como vigilância de vídeo, interação humano-computador e análise de comportamento humano. Além disso, espera-se que os resultados deste trabalho forneçam ideias úteis para o projeto de futuros sistemas de reconhecimento de ações que são sensíveis tanto à eficácia quanto à eficiência computacional.

1.4 Organização da Dissertação

Além deste capítulo introdutório, esta dissertação contém os seguintes capítulos:

Capítulo 2: Trata da revisão da literatura referente à estimação de poses humanas bidimensionais e tridimensionais;

Capítulo 3: Trata da revisão da literatura referente ao reconhecimento de ações humanas;

Capítulo 4: Apresenta a abordagem proposta neste trabalho para o reconhecimento de ações humanas baseada em poses 3D obtidas de poses 2D;

Capítulo 5: Apresenta o conjunto de dados utilizado neste trabalho para avaliação dos métodos propostos, detalha os experimentos realizados e discute os resultados obtidos, analisando o desempenho da abordagem proposta em comparação com abordagens de referência;

Capítulo 6: Expõe as principais conclusões do estudo, ressaltando suas contribuições, limitações e possibilidades para trabalhos futuros.

2 Estimação de Pose Humana

Este capítulo apresenta uma revisão da literatura sobre a estimação de poses humanas. Inicialmente, na Seção 2.1 o tema é introduzido, o conceito de pose humana é definido e as abordagens comumente adotadas pelos métodos para estimação de poses são apresentadas. Em seguida, na Seção 2.2, são apresentados métodos propostos para estimação de poses 2D. Após, na Seção 2.3, são apresentados métodos propostos para estimação de poses 3D. Finalmente, nas Seções 2.4 e 2.5, são apresentadas, respectivamente, uma abordagem multitarefa baseada em redes neurais convolucionais para realizar, de forma conjunta, a estimação de pose humana 3D e o reconhecimento de ações humanas a partir de imagens ou vídeos RGB, e uma abordagem para estimação de poses humanas 3D a partir de vídeo monocular RGB usando arquiteturas de *transformers* e GCN (Graph Convolutional Networks).

2.1 Introdução

A descrição de uma pose humana, conforme proposta por Stamou et al. (2005), compreende um corpo articulado, constituído por um conjunto de partes rígidas conectadas por juntas, permitindo movimentos translacionais e rotacionais em três graus de liberdade.

Nesse contexto, a estimação de pose humana torna-se um desafio que visa encontrar uma pose P dentro de um espaço Π , que engloba todas as poses articuladas. Em cenários de imagens RGB, a tarefa concentra-se na extração de um conjunto de características das imagens, correspondentes a cada junta do corpo humano.

Diversos modelos de representação de pose humana existem, como o cinemático (esqueleto humano), planar e volumétrico, cada um com propósitos específicos. O modelo cinemático destaca-se como o mais utilizado em tarefas de estimativa de poses 2D e 3D devido à sua flexibilidade e intuitividade (ZUFFI et al., 2012). No entanto, é importante ressaltar que esse modelo carece de informações relacionadas à forma e textura do corpo humano.

Inicialmente, as abordagens pioneiras para o problema de reconhecimento de poses se baseavam em técnicas de processamento de imagens tradicionais. Por exemplo, Yang and Ramanan (2013) propuseram uma abordagem que faz uso de histogramas de gradientes orientados (do inglês, *histogram of oriented gradients* - HOG) para extrair características que identificam cada junta do corpo. Contudo, essa metodologia mostrou-se limitada para obter poses precisas em ambientes do mundo real.

O desenvolvimento da área de aprendizado profundo, juntamente com o surgimento de redes convolucionais, revolucionou a detecção de poses a partir de imagens monoculares. Métodos baseados em aprendizado profundo têm apresentado resultados notáveis, superando

as abordagens tradicionais (DANG et al., 2019). O avanço nessa direção destacou-se pela capacidade desses métodos em lidar com desafios complexos na estimativa de poses.

Quanto às diferentes abordagens na estimação de poses por meio de redes em profundidade, Chen et al. (2020) propõem uma classificação considerando variações na modelagem do problema. Essas variações podem incluir a utilização ou não de poses para a obtenção de novas poses, a organização das redes neurais e a quantidade de pessoas na cena.

No tocante à utilização de poses, os métodos podem ser categorizados como generativos ou discriminativos. Abordagens generativas utilizam informações das juntas nos rótulos para derivar novas poses, enquanto métodos discriminativos buscam aprender uma função que mapeia um indivíduo em um espaço de poses, sem utilizar informações específicas dos modelos de poses. Esta função, por sua vez, pode escolher uma pose ou uma combinação de poses de um dicionário para representar a nova pose.

Outro critério de classificação é a arquitetura das redes neurais. Os métodos são divididos em abordagens de um único estágio, que empregam uma única arquitetura de ponta a ponta, e múltiplos estágios, que subdividem as tarefas entre diferentes redes neurais. Um exemplo de método de múltiplos estágios é a estimação de poses 3D com múltiplos indivíduos, onde três redes são definidas: uma para detectar os indivíduos, outra para encontrar suas poses 2D e, por fim, uma terceira para projetar as poses 2D em um espaço 3D. Em geral, as abordagens são divididas entre aquelas que tratam de uma única pessoa na imagem e aquelas que lidam com múltiplas pessoas.

Nas estratégias voltadas para uma única pessoa, a problemática está associada a situações em que há apenas um indivíduo na cena, delimitando claramente a região de interesse e simplificando a abordagem. No entanto, quando a cena envolve múltiplas pessoas, torna-se necessário realizar procedimentos prévios para isolar especificamente uma pessoa na imagem. Nesse contexto, destacam-se dois tipos de métodos amplamente empregados para lidar com essa complexidade: os métodos fundamentados em regressão e os métodos baseados em detecção.

- **Métodos baseados em regressão:** visam localizar todas as articulações do corpo em uma imagem através de redes ponta-a-ponta. No entanto, a abordagem inicial que considerava apenas as informações brutas das posições das articulações mostrou-se limitada ao negligenciar as informações nas proximidades das articulações. Para superar essa limitação, a supervisão foi ajustada, passando a utilizar mapas de calor que representam as possíveis vizinhanças das articulações. Um desafio subsequente surge ao utilizar exclusivamente mapas de calor, pois, dependendo da resolução, as informações relacionadas às articulações podem se tornar imprecisas no processo de decisão. Para mitigar esse problema, Nibali et al. (2018) propõem a introdução de uma camada de transformação numérica em redes neurais, permitindo o cálculo preciso das coordenadas das articulações a partir dos mapas de calor, especialmente em casos de baixa resolução.

- **Métodos baseados em detecção:** buscam identificar separadamente as partes do corpo para, em seguida, organizar essas partes de maneira a representar o corpo humano. Esses métodos tendem a ser mais sensíveis a variações na complexidade do fundo e à oclusão corporal. Para aprimorar o treinamento de redes convolucionais, Jain et al. (2015) propuseram a incorporação de um canal de mapas de calor associado a uma distribuição gaussiana centrada em cada junta, resultando em mapas de calor específicos para cada articulação. Essa abordagem tem se consolidado na detecção de partes do corpo. Além disso, Luvizon et al. (2019a) introduzem uma função do tipo *soft-argmax* para transformar redes neurais baseadas em detecção em redes de regressão, possibilitando a aplicação de redes ponta-a-ponta em ambientes de detecção.

A existência de múltiplos indivíduos em uma imagem pode significativamente ampliar a complexidade do desafio, requerendo a realização de pré-processamento para isolar as pessoas na imagem. Alternativamente, as arquiteturas podem ser ajustadas para identificar os indivíduos nas imagens; no entanto, mesmo assim, questões relacionadas à interação entre as pessoas ainda podem surgir.

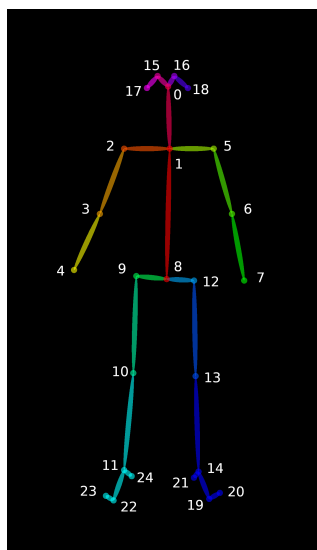
As abordagens para estimação de pose de múltiplas pessoas podem ser agrupadas em métodos *bottom-up* e *top-down*. Os métodos *top-down* utilizam um detector de pessoas e em seguida um detector de pose para cada uma das detecções da imagem, o que faz com que a complexidade do método seja proporcional ao número de pessoas detectadas. Os métodos *bottom-up* estimam a posição de cada junta do corpo humano na imagem e depois as associam para formar poses únicas, essa abordagem possui vantagens em relação ao tempo de execução, uma vez que não utiliza as informações globais da imagem de maneira direta (CAO et al., 2018).

Exemplos de métodos que utilizam a estratégia *bottom-up* são o *OpenPose* (CAO et al., 2018) e o *PifPaf* (KREISS et al., 2019). Enquanto no *OpenPose* é possível utilizar uma opção de formato de saída com um modelo formado por 25 pontos chaves (Figura 5a), incluindo cabeça, tronco e pés, o modelo do *PifPaf* conta com 17 pontos chaves (Figura 5b), e não possui opções para detectar informações dos pés.

2.2 Métodos de Estimação de Poses 2D

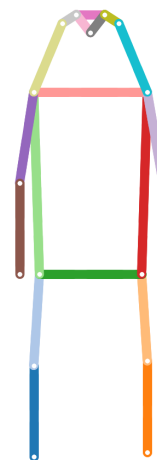
A estimação de poses 2D é uma área fundamental em visão computacional, com aplicações que vão desde o rastreamento de pessoas em vídeos até a análise de movimentos humanos. O objetivo dessas técnicas é inferir a posição e a orientação das articulações do corpo humano dentro de uma imagem ou um quadro de vídeo. Apesar do desafio imposto pelas variações de iluminação, oclusões e ambiguidades inerentes à projeção 2D, os avanços recentes, especialmente os baseados em métodos de aprendizado profundo, proporcionaram uma

Figura 5 – Modelos de poses 2D usados pelos métodos *OpenPose* (a) e *PifPaf* (b).



(a) Modelo do *OpenPose*.

Fonte: Cao et al. (2018).



(b) Modelo do *PifPaf*.

Fonte: Kreiss et al. (2019).

melhoria significativa na precisão e robustez da estimação de poses. Nesta seção, discutiremos os principais métodos de estimação de poses 2D.

2.2.1 *OpenPose*

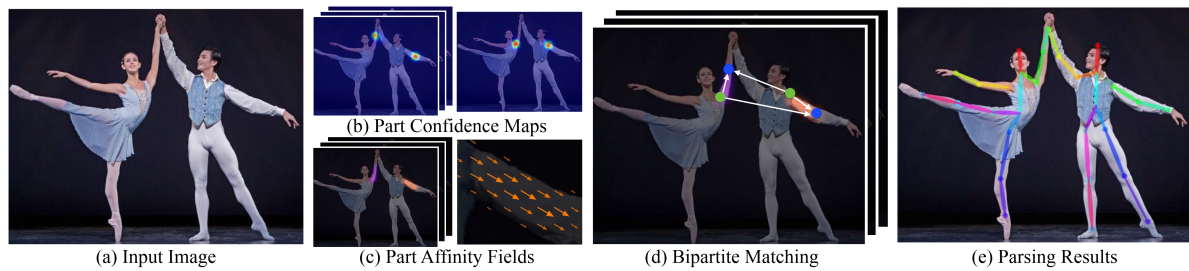
O OpenPose (CAO et al., 2018) é uma abordagem em tempo real para detecção de pose 2D de múltiplas pessoas em uma imagem, capaz de realizar as detecções com alta acurácia e um bom desempenho computacional. É o primeiro sistema de código aberto para detecção de pose 2D em tempo real que inclui corpo, pés, mãos e face. Diferente de muitas abordagens focadas em estimar a pose de indivíduos, essa aplicação trata a imagem inicialmente de forma global, com uma abordagem *bottom-up*, em contraste às abordagens anteriores que inicialmente detectavam cada indivíduo na imagem e suas respectivas poses um por vez, ocasionando uma grande aumento do esforço computacional em função da quantidade de pessoas na imagem.

O método ganha performance e mantém a acurácia ao utilizar uma abordagem diferente chamada de campos de afinidade, que mapeiam a posição e orientação de membros presentes no domínio da imagem através de um conjunto de vetores de duas dimensões, juntamente com os mapas de confiança, que representam a confiança de existir uma parte de corpo em determinado pixel da imagem. A partir dessa informação global codificada é possível adotar uma estratégia de detecção e associação "gulosa", capaz de produzir resultados precisos por um baixo custo computacional (CAO et al., 2018).

No processo de detecção de pose os campos de afinidade são melhorados iterativamente juntamente com os mapas de confiança por meio de duas redes neurais convolucionais (CNN) interligadas, uma para os campos de afinidade e outra para os mapas de confiança, que ao

trabalharem juntas melhoram o resultado iterativamente. Em seguida, as partes começam a ser associadas, de maneira que as associações com menor probabilidade de ocorrer e gerar poses não válidas vão sendo eliminadas iterativamente até atingir o resultado final, com os pontos conectados e as poses detectadas (CAO et al., 2018). A Figura 6 ilustra o *pipeline* do método.

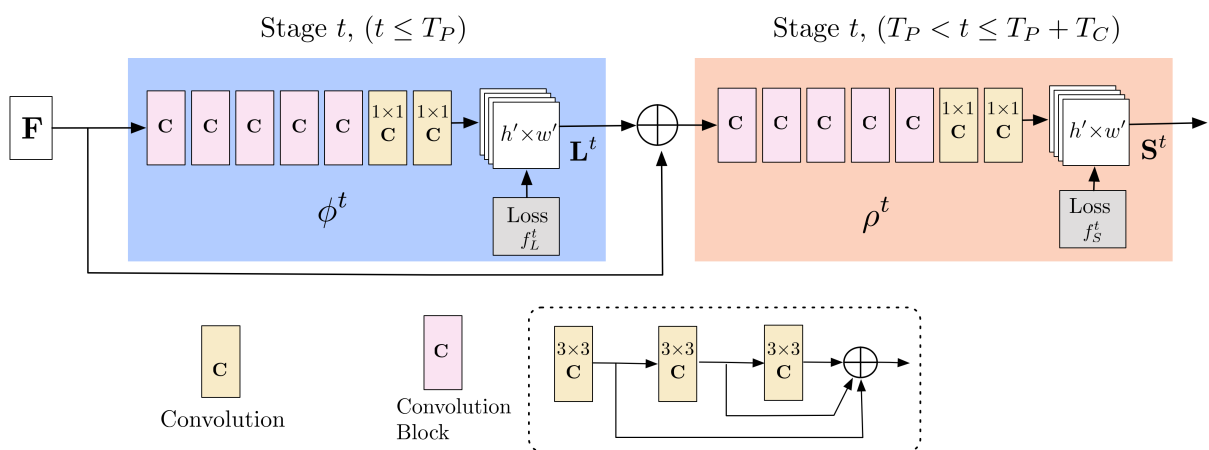
Figura 6 – *Pipeline* do método OpenPose: (a) entrada da imagem completa na CNN; (b) geração de mapas de confiança para detecção das partes do corpo; (c) geração de PAFs para associação das partes; (d) pareamento bipartido para associar candidatos a partes; (e) montagem das poses completas de todos os indivíduos



Fonte: Cao et al. (2018).

A Figura 7 mostra a arquitetura da CNN que prediz os campos de afinidade (que codificam a associação entre as partes) iterativamente, em azul e, em bege a detecção dos mapas de confiança. As previsões são refinadas iterativamente a cada estágio $t \in \{1, \dots, T\}$, com supervisão intermediária em cada estágio (CAO et al., 2018).

Figura 7 – Arquitetura da CNN multiestágio: (a) primeiros estágios preveem PAFs L_t ; (b) últimos estágios preveem mapas de confiança S_t ; (c) previsões e características da imagem de cada estágio são concatenadas para o estágio seguinte; (d) convoluções originais de *kernel* 7 são substituídas por três camadas de *kernel* 3 concatenadas.



Fonte: Cao et al. (2018).

2.2.2 PifPaf

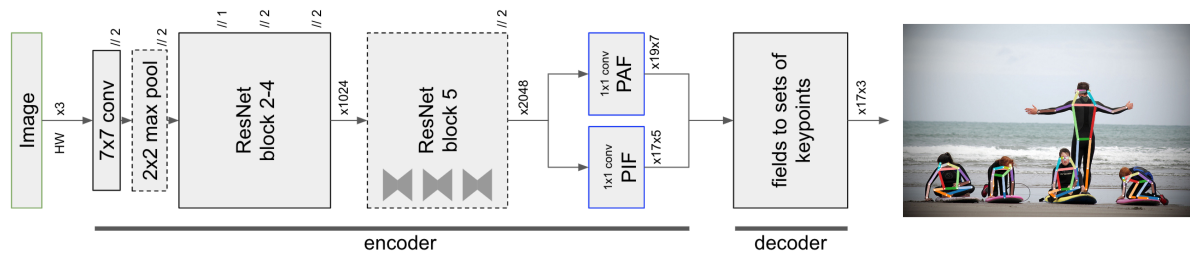
O PifPaf (KREISS et al., 2019) é um método de reconhecimento de pose 2D em tempo real capaz de reconhecer poses de múltiplas pessoas presentes em uma imagem. Este método

foi pensado para ser aplicado no ramo da navegação autônoma, tendo como principal foco a detecção de pessoas à longas distâncias, baixas resoluções e alto nível de oclusão com o objetivo de aumentar a segurança no trânsito de veículos autônomos e robôs de entrega.

Nesta abordagem, são usados dois tipos de rede neural para desempenhar dois papéis importantes na detecção das poses. Para cada parte de corpo presente na imagem uma rede prediz a confiança, a posição precisa e o tamanho. Essa informação é chamada de *Part Intensity Fields* (PIF) (KREISS et al., 2019) semelhante aos mapas de confiança utilizados no OpenPose (CAO et al., 2018).

A outra rede utilizada, que trata da associação entre as partes encontradas pelos campos de intensidade, é chamada de *Part Association Fields* (PAF). Graças à este passo do método, são obtidos resultados que superam as abordagens anteriores quando em imagens de baixa resolução e com diversas oclusões, pois para associar as partes e formar poses humanas completas os campos de associação de partes utilizam uma regressão baseada na função de perda de Laplace, que incorpora uma noção maior de incerteza (KREISS et al., 2019). A Figura 8 ilustra o *pipeline* do método, dada a imagem de entrada a rede neural codifica a imagem nos campos PIF e PAF e em seguida são decodificados formando as poses para cada pessoa na imagem, compostas por pares de coordenadas das juntas de cada pessoa acompanhadas de um grau de confiança (KREISS et al., 2019).

Figura 8 – Arquitetura do modelo: a entrada é uma imagem ($H \times W$) com três canais de cor ($\times 3$). O *encoder* da rede neural gera os campos PIF e PAF com 17 e 19 canais, respectivamente, e operações com *stride* 2 são indicadas por $//2$. O *decoder* converte os campos PIF e PAF em estimativas de pose com 17 articulações, cada uma representada por coordenadas x e y e um score de confiança.



Fonte: Kreiss et al. (2019).

2.3 Métodos de Estimação de Poses 3D

Atualmente já existem abordagens para a estimativa de poses 3D, mas essas soluções operam apenas em ambientes controlados, como as baseadas no sensor de profundidade Kinect ou na detecção por marcadores (CHEN et al., 2020). Isso impõe restrições significativas, destacando a necessidade de alternativas mais flexíveis para a estimativa de poses 3D. Essas alternativas devem ser aplicáveis em ambientes não controlados, preferencialmente utilizando imagens RGB monoculares, capturadas por câmeras acessíveis de baixo custo.

Uma abordagem inicial para resolver esse desafio envolveu a utilização de redes ponta-a-ponta para a regressão da pose em coordenadas tridimensionais (LI; CHAN, 2015). Contudo, essa abordagem é problemática devido à exigência de bases de dados capturadas em ambientes controlados para a estimação de poses 3D, o que limita a diversidade dos dados e prejudica a capacidade de generalização das redes.

Com o avanço e sucesso de técnicas de estimação de poses 2D, como os métodos OpenPose (CAO et al., 2018) e PifPaf (KREISS et al., 2019), estratégias que exploram a maturidade desses estimadores de poses 2D para a estimação de poses 3D têm ganhado popularidade. Essas estratégias adotam um processo de duas etapas: a primeira obtém poses 2D válidas, enquanto a segunda, a partir dessas poses 2D, estima as poses 3D.

A principal premissa é que poses em duas dimensões, cujos rótulos podem ser obtidos com facilidade em diferentes contextos, contêm informações suficientemente ricas para permitir a reconstrução da pose humana em três dimensões (MARTINEZ et al., 2017). No entanto, essa tarefa apresenta desafios significativos, pois a projeção de uma pose tridimensional em um plano bidimensional resulta na perda de informações de profundidade, tornando o processo de reconstrução ambíguo e não invertível. Isso implica que uma mesma pose em duas dimensões pode corresponder a diversas configurações possíveis no espaço tridimensional. Apesar dessa limitação, Martinez et al. (2017) demonstram que arquiteturas simples, baseadas em regressão direta de poses em duas para três dimensões, podem superar métodos mais complexos com treinamento de ponta a ponta em termos de acurácia. A partir deste ponto, diferentes estratégias têm sido propostas para lidar com essa ambiguidade e melhorar a estimativa da pose 3D. Nas seções a seguir, são discutidas abordagens representativas que exploram diferentes fontes de informação, como malhas corporais, dados temporais de vídeo, múltiplas visões e técnicas de adaptação entre domínios, com o objetivo de tornar o problema mais tratável e ampliar a aplicabilidade em cenários do mundo real.

2.3.1 *Human Mesh Recovery*

O método *Human Mesh Recovery* (HMR), proposto por Kanazawa et al. (2018), utiliza a estratégia de ponta a ponta para recuperar a forma 3D humana e a pose a partir de uma imagem de entrada, e possui duas etapas principais. Primeiro, a imagem é fornecida a uma rede neural convolucional (CNN) que foi pré-treinada para detecção de pontos chave 2D. A saída desta etapa é um conjunto de pontos 2D que representa a pose humana na imagem de entrada. Em seguida, esses pontos 2D são fornecidos a uma regressão baseada no modelo *Skinned Multi-Person Linear Model* (SMPL), desenvolvido por Loper et al. (2015), que mapeia a pose 2D para uma pose 3D. Este mapeamento é expresso matematicamente como:

$$S(P_{2d}) = P_{3d} \quad (1)$$

onde S é a função de regressão, P_{2d} é a pose 2D e P_{3d} é a pose 3D estimada.

O SMPL é um modelo de aprendizado de máquina popular para a representação do corpo humano em 3D que foi projetado para generalizar bem entre diferentes identidades individuais, tipos de corpos, e poses - uma característica essencial para tarefas como a análise de movimentos humanos. Ele define um corpo humano usando um veículo de malha 3D baseado em um conjunto de parâmetros de pose e forma. A forma é controlada por vetores de forma que representam a diversidade de corpos humanos através de fatores como peso e altura. A pose, por outro lado, é manipulada através de vetores de pose que codificam a orientação dos membros do corpo.

O HMR demonstrou ser eficaz em lidar com o problema da ambiguidade da profundidade por capturar a relação complexa e não-linear entre as poses 2D e os espaços de formas 3D. A rede é treinada de maneira supervisionada e não supervisionada, utilizando imagens em segundo plano para uma maior generalização.

Os experimentos demonstraram que o HMR é bem sucedido em inferir poses 3D complexas e realistas, bem como a forma do corpo a partir de imagens em 2D. Além disso, ao contrário de muitos métodos anteriores, o HMR vai além da estimativa de pontos de articulação e produz um modelo de corpo completo que é compatível com a malha de forma 3D. A Figura 9 ilustra alguns resultados obtidos pelo método HMR.

Figura 9 – Exemplos de resultados obtidos pelo método HMR.



Fonte: Adaptada de Kanazawa et al. (2018).

2.3.2 Video Inference for Human Body Pose and Shape Estimation

O método *Video Inference for Human Body Pose and Shape Estimation* (VIBE), proposto por Kocabas et al. (2020) para a estimativa da pose 3D e da forma do corpo humano, incorpora a informação temporal de vídeos na sua inferência, o que permite uma estimativa mais precisa e robusta, mitigando o problema do ruído nas poses 2D e as ambiguidades de profundidade.

Na primeira etapa do VIBE, uma CNN é usada para estimar a pose 2D do sujeito no vídeo. Em seguida, a informação temporal é incorporada por meio de uma rede neural recorrente

(RNN) que lida com as sequências de articulações 2D estimadas. Finalmente, semelhante ao HMR, uma função de regressão baseada no modelo SMPL é usada para mapear a pose 2D para uma pose 3D. Além de considerar as informações do *frame* atual, os autores incorporam a informação da pose no tempo $t - 1$ para a previsão da pose no tempo t , levando ao seguinte modelo:

$$S(P_{2d}^t, P_{3d}^{t-1}) = P_{3d}^t \quad (2)$$

onde S é a regressão, P_{2d}^t é a pose 2D no tempo t , P_{3d}^{t-1} é a pose 3D no tempo $t - 1$ e P_{3d}^t é a pose 3D estimada no tempo t .

Os experimentos mostraram que o método VIBE supera muitas estimativas baseadas em quadros únicos e é capaz de lidar consistentemente com vários indivíduos no vídeo. Além disso, a inclusão da informação temporal permite ao VIBE fazer previsões de movimento mais realistas e suaves em comparação com métodos que só consideram informações de quadros individuais. A Figura 10 mostra alguns resultados obtidos pelo HMR e pelo VIBE. É possível notar que o método VIBE apresenta resultados superiores, com maior precisão na estimação das poses.

Figura 10 – Comparação qualitativa dos resultados obtidos pelo VIBE (linha superior) e HMR (KANAZAWA et al., 2018) (linha inferior).



Fonte: Adaptada de Kocabas et al. (2020).

2.3.3 Aprendizado de Poses 3D a partir de Múltiplas Visões

Um método que se destaca pelo uso de poses 2D estimadas de diferentes ângulos para a formação da pose 3D, é o de Rhodin et al. (2018). Neste trabalho, os autores propõem um método de aprendizado não supervisionado para recuperar poses 3D a partir de múltiplas visões 2D. Eles treinaram uma rede neural para reconstruir a pose 3D do sujeito a partir de várias visões 2D e, ao mesmo tempo, re-projetar a pose 3D de volta em 2D para todas as visões de entrada. Este método permite o reconhecimento de ações em cenários onde o sujeito é parcialmente ocluído em uma visão, pois as outras visões podem fornecer informações perdidas. A ideia do método, de se estimar poses 3D a partir de diferentes observações 2D, é ilustrada pela Figura 11.

Figura 11 – Estimação da pose 3D a partir de diferentes ângulos de visão 2D.



Fonte: Adaptada de Rhodin et al. (2018).

Para atingir esse objetivo, os autores propõem o uso de um modelo de rede neural profunda treinado para prever poses 3D a partir de imagens multi-visão. Especificamente, eles utilizam as multi-visões para fornecer supervisão fraca e forçar o sistema a prever a mesma pose em todas as visões, conforme ilustrado na Figura 11. A consistência entre as visões força o sistema a prever poses consistentes em diferentes pontos de vista. Além disso, o método propõe o uso de uma distância de pose normalizada (NPD) utilizada como função de perda na estimativa das poses. O cálculo dessa distância é dado pela Equação 3, onde N é o número de pontos de pose, p_i é o ponto de pose real, $\hat{p}_i$ é o ponto de pose previsto pelo modelo e s_i é o vetor de escala correspondente ao ponto de pose.

$$NPD = \frac{1}{N} \sum_{i=1}^N \frac{\|p_i - \hat{p}_i\|^2}{\|s_i\|} \quad (3)$$

A utilização dessa distância de pose normalizada é crucial para manter a precisão quando os dados anotados são limitados (RHODIN et al., 2018).

O processo de treinamento do modelo envolve a utilização de um pequeno conjunto de imagens com poses verdadeiras, as quais servem tanto para fornecer supervisão forte durante o treinamento quanto para regularizar as previsões de multi-visão, fazendo com que elas permaneçam próximas das previsões de uma rede treinada apenas com os escassos dados supervisionados. Esse processo é crucial para evitar que o modelo aprenda a sempre prever a mesma pose independentemente da entrada. Os resultados apresentados demonstraram a eficácia da estratégia de treinamento multi-visão fracamente supervisionada, reduzindo significativamente a necessidade de dados anotados e aumentando a robustez do algoritmo em relação a mudanças de ponto de vista e escala (RHODIN et al., 2018).

2.3.4 Aprendizado de Poses 3D a partir de Imagens Monoculares

O método Domain-Unified Approach (DUA), proposto por Manesco et al. (2023), implementa uma estratégia em duas etapas para lidar com o problema mal posto de inferência de pose 3D a partir de uma única imagem 2D, oferecendo uma solução mais tratável que não depende fortemente de grandes quantidades de dados pareados em 3D. Um problema crítico em estimadores de pose 3D monocular é a degradação de desempenho causada pela ausência de dados que representem configurações de pose não observadas (MARTINEZ et al., 2017), o que afeta a qualidade dos esqueletos preditos. O DUA aborda essa questão integrando três componentes ao processo de aprendizado: adaptação de domínio a partir de dados sintéticos para aumentar a diversidade do treinamento, uma representação unificada de pose para lidar com diferentes modelos esqueléticos e uma metodologia de treinamento específica voltada para mitigar a propagação de erro em articulações com poses extremas.

O DUA realiza estimação de pose entre domínios por meio de otimização adversarial entre dois objetivos: minimizar o erro de reconstrução e maximizar a discrepância de características entre domínios. A arquitetura é inspirada em redes neurais adversariais de domínio (*Domain-Adversarial Neural Networks*, DANN) (GANIN et al., 2016), com modificações focadas na estimação de pose 3D com base em trabalhos anteriores (MARTINEZ et al., 2017). Para um estimador de pose Π e uma imagem de entrada x , a função de perda composta combina regularização L_1 e L_2 :

$$\mathcal{L}_{pose}(x) = \beta(y - \Pi(x))^2 + (1 - \beta)\|y - \Pi(x)\|, \quad (4)$$

onde $0 \leq \beta \leq 1$ é um hiperparâmetro que controla a importância de cada parte da perda de pose. O componente adversarial treina um discriminador de domínio G para distinguir características de pose sintéticas das reais, enquanto Π aprende a gerar representações invariantes ao domínio por meio de reversão de gradiente.

Representações de esqueletos incompatíveis entre diferentes conjuntos de dados (por exemplo, hierarquias de juntas do SMPL vs. Human3.6M) criam desafios de compatibilidade para o aprendizado entre domínios, um problema comumente relatado na literatura (SÁRÁNDI et al., 2023; RAPCZYŃSKI et al., 2021). Embora essa incompatibilidade possa ser facilmente resolvida em alguns casos (quando dois conjuntos compartilham um subconjunto comum de articulações), ao lidar com conjuntos de dados de pose humana 3D obtidos por sensores de captura corporal distintos, a posição geral das juntas pode estar incorreta, levando a uma relação antropomórfica inválida entre as articulações.

Diferentemente de outras abordagens presentes na literatura, que frequentemente utilizam normalizações entre juntas, o DUA busca encontrar uma representação comum entre diferentes entradas de pose por meio de uma rede neural de conversão $\varphi : \mathcal{A} \mapsto (\mathcal{B} - \mathcal{A})$, que resolve esse problema utilizando deslocamentos aprendidos. Dado um conjunto de articulações

de entrada $X_{\mathcal{A}}$ na representação $\mathcal{A}$, o conversor prediz vetores de deslocamento:

$$X_{\mathcal{B}} = X_{\mathcal{A}} + \varphi(X_{\mathcal{A}}; \theta_{\varphi}) \quad (5)$$

onde θ_{φ} representa os parâmetros da rede neural. Essa formulação residual preserva as proporções corporais intrínsecas ao modificar posições relativas das articulações, ao invés de coordenadas absolutas. Essencialmente, a rede neural φ observa a pose de entrada $X_{\mathcal{A}}$ e calcula o quanto cada articulação precisa se mover no espaço tridimensional para reconstruir a pose equivalente no formato $\mathcal{B}$. Em vez de substituir completamente as posições originais das articulações, as poses são ajustadas por meio desses deslocamentos aprendidos.

Estimativa de Incerteza

A estimação de pose humana 3D a partir de vídeos monoculares enfrenta desafios inerentes relacionados à propagação de erro, especialmente para articulações em posições cinemáticas extremas. Esse problema é agravado pela natureza mal colocada da tarefa de inferir estrutura tridimensional a partir de uma única visão, resultando em ambiguidades causadas por auto-occlusão e variações na posição da câmera. Para lidar com isso, o método DUA incorpora um mecanismo para quantificar e mitigar as incertezas provenientes desses cenários desafiadores.

No contexto do aprendizado profundo Bayesiano, a incerteza geralmente é categorizada em duas formas principais: a incerteza epistêmica, que reflete a falta de conhecimento do modelo e pode ser reduzida com mais dados de treinamento, e a incerteza aleatória, que modela a variabilidade inerente ou o ruído presente nas observações em si (KENDALL et al., 2018).

O método DUA segue uma abordagem ingênua para quantificação de incerteza, utilizando características extraídas do estimador de pose subjacente para prever a probabilidade de uma estimativa incorreta para cada articulação. Uma variável aleatória $\mathcal{U}_j$ é associada a cada articulação j , representando sua incerteza estimada. Essa variável é gerada a partir do mapeamento de uma distância Euclidiana normalizada, Δ_j , que quantifica a discrepância da articulação j . Articulações com valores baixos de discrepância ($\Delta_j \approx 0$) são mapeadas para $\mathcal{U}_j \approx 0$ (baixa incerteza), enquanto aquelas com discrepâncias significativas são mapeadas para $\mathcal{U}_j \approx 1$ (alta incerteza). Esse mapeamento auxilia na avaliação e quantificação da incerteza associada a previsões individuais de articulações, reduzindo a incerteza global em grupos cinemáticos extremos.

2.4 Método Híbrido para Estimação de Poses 3D e Reconhecimento de Ações

Luvizon et al. (2019b) propõem uma abordagem multi-tarefa baseada em redes neurais convolucionais para realizar, de forma conjunta, a estimação de pose humana em 3D e o reconhecimento de ações humanas a partir de imagens ou vídeos RGB. A arquitetura proposta permite que ambas as tarefas sejam treinadas de forma simultânea em uma única rede, compartilhando grande parte das camadas convolucionais, o que resulta em uma solução eficiente e com alto desempenho. A entrada do modelo pode ser uma imagem estática $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ ou uma sequência de vídeo $\mathbf{V} \in \mathbb{R}^{T \times H \times W \times 3}$, onde T é o número de quadros e $H \times W$ são as dimensões da imagem. As saídas incluem a pose humana estimada $\hat{\mathbf{p}} \in \mathbb{R}^{N_j \times 3}$ e o vetor de probabilidades para as ações $\hat{\mathbf{a}} \in \mathbb{R}^{N_a}$.

A rede é composta por um fluxo inicial de convoluções (entry-flow) seguido de blocos de predição organizados em pirâmides de escalas, intercaladas com unidades de redução (downscaling) e ampliação (upscaling) de resolução. Cada bloco de predição gera mapas de probabilidade para as articulações $h_t^{p,l} \in \mathbb{R}^{H_f \times W_f \times N_j}$ e mapas de profundidade $d_t^{p,l} \in \mathbb{R}^{H_f \times W_f \times N_j}$, os quais são utilizados para estimar a pose 3D por meio de uma abordagem diferenciável. A estimação das coordenadas 2D de cada junta j é realizada através da operação soft-argmax (LUVIZON et al., 2019b), definida como:

$$\Psi(h_j) = \left(\sum_{c=0}^{W_h} \sum_{r=0}^{H_h} \frac{c}{W_h} \cdot h_{r,c}, \sum_{c=0}^{W_h} \sum_{r=0}^{H_h} \frac{r}{H_h} \cdot h_{r,c} \right), \quad (6)$$

onde h_j é o mapa de probabilidade da junta j , com dimensões $H_h \times W_h$.

A profundidade de cada junta z_j é obtida por meio de uma média ponderada do mapa de profundidade com o mapa de probabilidade correspondente:

$$z_j = \sum_{c=0}^{W_h} \sum_{r=0}^{H_h} h_{r,c}^j \cdot d_{r,c}^j, \quad (7)$$

onde $h_{r,c}^j$ e $d_{r,c}^j$ são, respectivamente, os valores dos mapas de probabilidade e profundidade para a junta j . A concatenação das estimativas (x_j, y_j, z_j) resulta na pose tridimensional $\hat{p}_j$ da junta j . As estimativas de pose são reinjetadas na rede para refinar predições em blocos subsequentes, conforme sugerido na Equação 8:

$$X_t^{p,l} = W_r^{p,l} * h_t^{p,l} + W_s^{p,l} * d_t^{p,l} + Z_t^{p,l} + Z_t^{p,l}. \quad (8)$$

Para o reconhecimento de ações, a rede explora duas fontes principais de informação: características visuais extraídas diretamente das imagens e as poses estimadas. As poses ao longo do tempo são organizadas em uma representação similar a uma imagem, onde o tempo

corresponde à dimensão vertical, as articulações à horizontal, e as coordenadas x , y , z são codificadas como canais. As características visuais são extraídas ponderando os mapas de ativação da rede com os mapas de probabilidade das articulações, o que permite focar em regiões relevantes da imagem. Ambas as representações são combinadas e refinadas ao longo da rede. Para evitar que os gradientes das duas tarefas interfiram negativamente entre si, os autores propõem a duplicação das camadas de regressão de pose no final do treinamento, desacoplando as poses utilizadas para a tarefa de ação daquelas supervisionadas diretamente, conforme demonstrado por seus experimentos.

2.5 Transformer-GCN com Aprendizado de Representações Contextualizadas

Ye et al. (2024) propõem uma abordagem inovadora para estimação de poses humanas em 3D a partir de vídeo monocular RGB, combinando os pontos fortes das arquiteturas *Transformer* e *Graph Convolutional Networks* (GCNs) em um modelo de fluxo duplo. O método busca superar limitações típicas de abordagens monoculares, como a ambiguidade de profundidade e a escassez de dados anotados em 3D, por meio de uma estratégia de pré-treinamento baseada em aprendizado de representações contextuais.

A arquitetura é composta por dois fluxos paralelos: o fluxo GCN, responsável por capturar relações locais entre articulações e quadros adjacentes, e o fluxo Transformer, que modela relações espaciais e temporais de longo alcance. As saídas desses dois módulos são fundidas de maneira adaptativa, permitindo ao modelo combinar de forma equilibrada as dependências locais e globais.

No estágio de pré-treinamento, é utilizada uma estratégia de auto-destilação inspirada em técnicas como MAE e Data2Vec, em que partes das entradas são mascaradas e o modelo é treinado para reconstruir representações contextuais a partir dessas entradas parciais. Este treinamento não supervisionado permite aproveitar grandes volumes de dados com anotações apenas em 2D. O modelo final é ajustado para prever poses 3D diretamente a partir dessas representações enriquecidas.

Para avaliar a capacidade do modelo proposto no reconhecimento de ações humanas, foram conduzidos experimentos com o conjunto de dados NTU RGB+D, utilizando duas divisões principais dos dados: *Cross-Subject* (X-Sub), onde diferentes sujeitos são usados para treinamento e teste, e *Cross-View* (X-View), onde diferentes ângulos de câmera são utilizados.

A Tabela 1 apresenta os resultados obtidos utilizando o modelo Transformer-GCN proposto, em duas configurações: base (B) e ampliada (L), com e sem pré-treinamento. Como métrica de avaliação, foi utilizada a acurácia Top-1.

Os resultados demonstram que o modelo proposto superou o desempenho do Moti-

Tabela 1 – Acurácia Top-1 no conjunto de dados NTU RGB+D.

Método	NTU-RGB (<i>X-Sub</i>)	NTU-RGB (<i>X-View</i>)
<i>MotionBERT</i> (ZHENG et al., 2022a)	87.6%	93.4%
Modelo-B (sem pré-treinamento)	87.0%	91.9%
Modelo-B (com pré-treinamento)	88.1%	93.4%
Modelo-L (sem pré-treinamento)	88.0%	93.0%
Modelo-L (com pré-treinamento)	89.5%	93.5%

onBERT (ZHENG et al., 2022a) na divisão *X-Sub*, alcançando uma acurácia de 89.5% na versão ampliada com pré-treinamento, o que representa um ganho de 1.9 pontos percentuais. Já na divisão *X-View*, os métodos obtiveram desempenho semelhante, com leve vantagem para o modelo proposto. Isso valida a eficácia da estratégia de pré-treinamento e a arquitetura *dual-stream* baseada em Transformer e GCN para a tarefa de reconhecimento de ações humanas com poses 3D.

3 Reconhecimento de Ações Humanas

Este capítulo apresenta uma revisão da literatura sobre o reconhecimento de ações humanas. Inicialmente, na Seção 3.1 o tema é introduzido e o conceito de ação humana é definido. Em seguida, na Seção 3.2, são exemplificadas aplicações do mundo real que podem ser beneficiadas sobremaneira com o reconhecimento automático de ações humanas. Após, na Seção 3.3, são apresentados os desafios mais importantes das pesquisas nesta área do conhecimento. Finalmente, na Seção 3.4, são apresentados métodos estado da arte propostos para o reconhecimento de ações humanas.

3.1 Introdução

Reconhecimento pode ser definido como a tarefa em que se busca determinar se um dado de entrada possui ou se assemelha a algum objeto, característica ou atividade específica (AHAD, 2011). Na Visão Computacional, o reconhecimento de ações humanas busca compreender e classificar componentes de ações ou atividades em uma cena contida em uma imagem ou em um vídeo. Essa é uma tarefa muito desafiadora, uma vez que as ações possuem um caráter dinâmico e usualmente são muito difíceis de serem reconhecidas a partir da análise de momentos únicos no tempo (AHAD, 2011). Além disso, um nível a mais de complexidade é adicionado à equação, uma vez que existem casos que apresentam uma grande variação intra-classe, nos quais, por exemplo, uma mesma ação pode ser executada de maneiras diferentes por cada pessoa, e casos que apresentam uma grande semelhança inter-classe, como correr e andar, que, apesar de serem muito semelhantes, caracterizam ações distintas (KONG; FU, 2022).

Não existe um consenso na literatura a respeito da definição formal do que é uma ação humana. Esse termo é utilizado para definir desde simples movimentos de membros individuais, como chutar, por exemplo, até movimentos complexos de todo o corpo humano, podendo até ser combinado com interações com outros atores ou objetos da cena (KONG; FU, 2022). No entanto, de acordo com Camarena et al. (2023), podemos dizer de uma maneira generalista que a ação humana é uma entidade que pode ser observada por outra entidade, como um outro ser humano ou um computador, e essa entidade observadora é capaz de decodificar a informação (gestos) observada através de sensores, como olhos humanos ou câmeras, por exemplo. A partir dessas observações, uma pessoa é capaz de reconhecer com naturalidade o que outras estão fazendo, no entanto, essa não é uma tarefa simples para um computador.

3.2 Aplicações no Mundo Real

Apesar de ser estudado já há décadas, o reconhecimento de ações humanas vem se mantendo como um assunto em destaque no contexto de visão computacional, uma vez que seu uso oferece grande valor para diversas aplicações em diversas áreas (SHAIKH et al., 2024). Dentre as diversas aplicações do reconhecimento de ações humanas, Ahad (2011) destaca:

Segurança: Nos dias atuais, a segurança vem se tornando uma preocupação cada vez maior no dia-a-dia da vida urbana. Em diversos locais podem ser encontradas câmeras de vídeo utilizadas para o monitoramento de pessoas e veículos a fim de se garantir o cumprimento das regras e leis de segurança nestes locais (KONG; FU, 2022). A utilização de algoritmos para reconhecimento de ações em imagens capturadas pelos sistemas de vigilância pode contribuir sobremaneira para o aumento da segurança nesses locais, uma vez que ações suspeitas ou indesejadas podem ser detectadas em tempo real, mitigando os riscos causados por ações criminosas ou que transgridam as regras vigentes (PATRIKAR; PARATE, 2022);

Recuperação de Informação: Com o crescimento acelerado da quantidade de informações disponíveis na *Internet*, métodos para recuperação da informação cada vez mais sofisticados se fazem necessários. Além de dados em formato de texto, atualmente existe uma grande necessidade de gerenciar e recuperar imagens e vídeos de maneira eficiente, pois os motores de busca mais comuns utilizam informações textuais, tais como título, descrição e palavras chaves usadas para indexar os dados multimídia, que por sua vez podem estar incorretas ou serem irrelevantes em relação ao conteúdo do vídeo ou da imagem (KONG; FU, 2022). Neste contexto, o reconhecimento de ações poderia ser aplicado aos vídeos, de forma a extrair as características contidas no material, como foi feito no trabalho de Ciptadi et al. (2014), onde os autores criaram um *framework* de recuperação de vídeos baseado na similaridade entre as representações de ações extraídas das imagens;

Interação Humano-Máquina: Técnicas de interação entre humano e robô são comumente aplicadas a ambientes de indústria. Esse tipo de aplicação visa tornar mais fluida a interação entre um humano e uma máquina, muitas vezes com o intuito de trabalharem em conjunto para alcançar um objetivo comum. Nesse contexto, é importante que o robô entenda as ações e a intenção do trabalhador humano, a fim de responder aos gestos e ações de maneira adequada. Para isso, métodos de reconhecimento de ações podem ser aplicados (ZHENG et al., 2022b). Além disso, em contextos mais gerais, o reconhecimento de ações pode ser utilizado no treinamento de robôs que executam tarefas humanas, funcionando como uma forma de codificar as ações que se deseja executar (KONG; FU, 2022);

Navegação de Carros Autônomos: No contexto dos carros autônomos, é essencial para os sistemas de navegação a possibilidade de perceber as intenções e prever situações não só envolvendo outros veículos, mas também os pedestres. Os algoritmos de predição de ações buscam prever a intenção de uma pessoa em um curto período de tempo, durante o início da ação. Neste contexto, a predição difere do reconhecimento de ações, uma vez que o mesmo busca identificar a ação após ela ocorrer completamente, no entanto, ambos possuem a mesma premissa, na qual se busca compreender as ações humanas com base na composição de movimentos do corpo humano (KONG; FU, 2022). Em uma situação de emergência, um veículo equipado com essa funcionalidade poderia analisar o movimento do corpo humano, estimar a possível ação enquanto a mesma ainda ocorre, evitando possíveis acidentes no trânsito.

3.3 Desafios

A capacidade desenvolvida de forma natural entre os seres humanos de entender ações e intenções de outras pessoas, é uma importante habilidade social para o convívio em sociedade (KONG; FU, 2022). No entanto, computadores enfrentam diversos desafios ao realizar o reconhecimento dessas ações, que de acordo com Camarena et al. (2023), podem ser divididos em cinco categorias: problemas relacionados à variabilidade entre as classes de ações, capacidade dos sensores, condições do ambiente, restrições no acesso à conjuntos de dados e as limitações computacionais. A partir da análise e compreensão de cada um desses desafios, é possível propor estratégias para superá-los, a fim de aprimorar o desempenho dos métodos para reconhecimento de ações.

3.3.1 Variabilidade Inter e Intra-classe

No reconhecimento de ações humanas, a variabilidade intra-classe pode ser definida como as diferenças existentes entre múltiplas execuções de uma mesma ação. Essas variações podem ocorrer devido à diversos fatores como as características antropométricas do indivíduo, velocidade de execução da ação, proporções corporais ou idade (CAMARENA et al., 2023). Além das variações em relação a diferentes indivíduos executando a mesma ação, uma mesma pessoa pode executar uma ação repetidas vezes, e dificilmente todas serão executadas exatamente da mesma maneira.

Por outro lado, a variabilidade inter-classes se refere à similaridades existentes entre ações diferentes. Como exemplo, podemos considerar o ato de se sentar e de se levantar. Apesar de serem ações opostas, elas apresentam a mesma característica estrutural e semântica, apresentando um caso muito desafiador para modelos que não levam em consideração a informação temporal. Outro exemplo é o caso das ações de caminhar e correr, que podem ser analisadas como variações de uma mesma classe de ação (CAMARENA et al., 2023).

3.3.2 Capacidade dos Sensores

Os sensores são uma peça muito importante para sistemas de visão computacional, uma vez que têm a função de capturar os dados que serão utilizados pela aplicação, sendo as câmeras RGB o sensor mais comumente utilizado para sistemas desse escopo. Nesse contexto, as câmeras apresentam algumas limitações que podem dificultar o desempenho da aplicação, como o campo de visão limitado, o que ocasiona oclusões, seja por conta de sobreposição de objetos, condições de iluminação ou simplesmente o posicionamento do sujeito no plano na câmera (CAMARENA et al., 2023). A capacidade de lidar com informações incompletas ou faltantes é uma característica essencial para o bom funcionamento de um sistema de reconhecimento de ações humanas em ambientes de mundo real.

Além disso, as câmeras RGB comuns apresentam problemas quanto à captura de informações de profundidade, e as noções de velocidade de objetos podem sofrer distorções de acordo com sua distância em relação à câmera. Outra questão que deve ser levada em consideração é a qualidade de captura das imagens. Imagens em resoluções muito baixas podem resultar na perda de informações, principalmente para objetos mais distantes (CAMARENA et al., 2023).

Além das câmeras RGB, sensores que capturam informações de profundidade, como os sensores RGB-D (ex.: Kinect), têm se destacado como uma alternativa viável para estimar poses humanas em 3D. Tais sensores combinam imagem colorida com dados de profundidade por meio de projeções infravermelhas, permitindo a reconstrução do esqueleto humano em três dimensões sem a necessidade de marcadores corporais ou múltiplas câmeras. Entre suas principais vantagens estão o baixo custo, a facilidade de uso, e a capacidade de operar com apenas um sensor, o que torna esses dispositivos atrativos para aplicações práticas. No entanto, apresentam limitações importantes, como o alcance reduzido (aproximadamente de 0,8 até 4 metros) e a sensibilidade à luz solar, o que restringe seu uso a ambientes internos. Além disso, a oclusão de partes do corpo causada pelo próprio sujeito ou por obstáculos, pode gerar erros significativos na estimativa das posições articulares, o que afeta diretamente a acurácia da pose 3D. Ainda assim, estudos mostram que, mesmo com imprecisões, sensores como o Kinect são capazes de capturar com sucesso os padrões de movimento e alcançar bons resultados em tarefas de reconhecimento de ações humanas (HAN et al., 2013).

3.3.3 Condições Ambientais

As condições ambientais podem ser um fator decisivo para o bom funcionamento de uma aplicação de reconhecimento de ações no mundo real, podendo influenciar diretamente na acurácia do modelo. Diversos fatores podem influenciar nesse caso, como condições climáticas desfavoráveis quando se trata de ambientes externos, quantidades e interações entre pessoas e objetos gerando planos de fundo dinâmicos e imprevisíveis, o que aumenta a complexidade da tarefa de diferenciar o que deve ser analisado como elemento principal e o que deve ser

desconsiderado pelo modelo. Além desses fatores, a iluminação também é crucial. Muitas sombras ou até mesmo a super-exposição de luz podem gerar muitos ruídos ou impossibilitar o reconhecimento das ações (CAMARENA et al., 2023).

3.3.4 Acesso à Conjuntos de Dados

O desempenho de um modelo de aprendizado de máquina depende de um treinamento adequado, com dados que representem as amostras que serão coletadas durante a execução em ambientes reais em nível de complexidade e diversidade. No entanto, a utilização de um conjunto de dados mais próximo possível do ideal pode demandar uma grande quantidade de tempo e esforço. De acordo com Camarena et al. (2023) uma alternativa, é a construção de um conjunto de dados totalmente novo, tomando como base os casos de uso específicos do modelo que será construído. Apesar de representar muito bem as amostras que serão encontradas em um cenário produtivo, o processo de criação é muito trabalhoso, demandando diversas etapas como a coleta, rotulagem de dados e pré-processamento, o que pode levar meses para ser concluído.

Outra possibilidade, é o uso de bases de dados já conhecidas na literatura, com métodos de avaliação pré-definidos, o que permite uma avaliação de resultados comparável à métodos estado-da-arte através de uma revisão de literatura. No entanto, pode ser desafiador encontrar a base de dados que se adeque ao caso de uso de maneira satisfatória, pois existe uma disponibilidade limitada de dados rotulados, que podem não ser suficientes para o treinamento de uma arquitetura profunda, além do fato das classes de ações serem pré-definidas e poderem não se encaixar no domínio alvo da aplicação (CAMARENA et al., 2023).

3.3.5 Limitações Computacionais

Por fim, um ponto importante de ser levado em consideração, em contraste com a quantidade de dados utilizada no treinamento, bem como a complexidade do modelo, são as limitações computacionais impostas. Quanto maior o conjunto de dados no treinamento, melhor o desempenho na classificação, no entanto, existem modelos que podem demandar uma quantidade massiva de dados, que resultaria em um treinamento muito demorado, demandando uma capacidade de computação que pode não estar disponível (CAMARENA et al., 2023). Além disso, se o modelo apresentar uma complexidade muito alta, o mesmo pode se tornar inutilizável em ambientes reais, principalmente se tratando de aplicações em tempo real.

3.4 Métodos para Reconhecimento de Ações Humanas

No contexto atual da Ciência da Computação, o reconhecimento de ações desempenha um papel crucial em diversas aplicações, como visão computacional, interação humano-máquina e análise de vídeos. O advento de tecnologias avançadas, como redes convolucionais de grafos

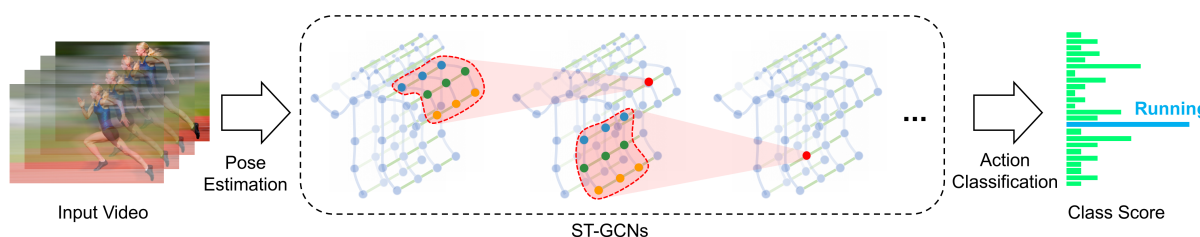
e *transformers*, tem revolucionado significativamente o campo, elevando o desempenho e a precisão dos métodos de reconhecimento de ações a níveis até então considerados estado-da-arte. Esta seção busca explorar e apresentar em detalhes esses métodos inovadores, destacando suas vantagens, desafios e contribuições para a vanguarda da pesquisa nesse domínio

3.4.1 Modelo Baseado em Grafo Espaço-Temporal

O método proposto por Yan et al. (2018) apresenta uma solução que utiliza as propriedades intrínsecas das redes neurais convolucionais (do inglês, *convolutional neural network*, CNN) relacionadas à representação local e hierárquica de informações em uma abordagem de redes convolucionais de grafos (do inglês, *graph convolutional network*, GCN), unindo informações espaciais e temporais graças à sua modelagem de grafo espaço-temporal. Neste modelo, os esqueletos são obtidos por meio de dispositivos de captura de movimentos ou estimadores de poses em vídeos. O resultado obtido por esses métodos são geralmente uma sequência de coordenadas 2D ou 3D das juntas do corpo humano em cada quadro do vídeo. Essas coordenadas são a base para construir o grafo espaço-temporal.

A construção do grafo espaço-temporal se dá em duas etapas, inicialmente, utiliza-se as informações espaciais em cada quadro do vídeo para conectar os nós dos grafos (juntas) utilizando as conexões físicas, ou seja, os membros. Em seguida cada grafo subsequente é conectado com o seu vizinho de acordo com a sequência dos *frames*, formando as conexões temporais. Dessa forma, a entrada da rede espaço-temporal convolucional de grafo (do inglês, *spatio-temporal graph convolutional network*, ST-GCN) (YAN et al., 2018) são os vetores de coordenadas dos nós do grafo, semelhante às CNNs baseadas em imagens, onde a entrada é uma matriz 2D com o valor de intensidade dos pixels da imagem. Na Figura 12 é possível visualizar essa analogia. Em seguida, o dado de entrada passa por múltiplas camadas convolucionais de ST-GCN, que resultam em mapas de características de alto nível, que por fim são passados como entrada por um classificador *SoftMax*, obtendo como resposta a classe à qual a ação pertence (YAN et al., 2018).

Figura 12 – Fluxo do método ST-GCN: estimacão de poses em vídeo, construção do grafo espaço-temporal e classificacão das açoes via *Softmax* (YAN et al., 2018).



Fonte: Adaptada de Yan et al. (2018).

Construção do Grafo

Técnicas anteriormente propostas, baseadas na análise de sequência de esqueletos normalmente representados por coordenadas 2D ou 3D de cada junta do corpo humano para cada quadro do vídeo, utilizaram um vetor único, concatenando as informações de todas as juntas para cada *frame* (KIM; REITER, 2017). Por outro lado, ao utilizar uma representação de grafo espaço-temporal para modelar as poses, Yan et al. (2018) representam as sequências de poses de maneira hierarquizada, onde o grafo espaço-temporal não direcionado é representado como $G = (V, E)$, em uma sequência de T poses contendo N juntas cada.

O conjunto de nós $V = \{v_{ti} | t = 1, \dots, T, i = 1, \dots, N\}$ inclui todas as juntas do esqueleto do indivíduo em todos os quadros do vídeo. Dessa forma, a entrada da ST-GCN em um nó específico é um vetor de características $F(v_{ti})$ que consiste em um vetor com as coordenadas da i -ésima junta no t -ésimo *frame* do vídeo de entrada, e o seu respectivo índice de confiança proveniente da detecção de poses.

O conjunto de arestas E é composto por dois subconjuntos E_S e E_F , um contendo as conexões de arestas espaciais e outro contendo as conexões temporais, respectivamente. O conjunto $E_S = \{v_{ti}v_{tj} | (i, j) \in H\}$ representa as conexões naturais existentes entre as juntas do corpo do indivíduo, interligadas fisicamente pelos seus membros, ou seja, a aresta entre o i -ésimo e o j -ésimo nó do grafo, em um mesmo *frame* t , onde H é o conjunto de conexões existentes no corpo humano (YAN et al., 2018). Já o conjunto $E_F = \{v_{ti}v_{t+1}i\}$ representa as conexões temporais entre o mesmo nó i em *frames* subsequentes t e $t + 1$. Dessa maneira, todas as arestas de um nó i representarão sua trajetória através do tempo.

CNN Baseada em Grafo Espacial

A GCN pode ser entendida como uma adaptação das redes neurais convolucionais para imagens, aplicada no contexto de grafos. Sendo assim, dois principais conceitos das CNNs de imagens são adaptados nas GCNs: a função de amostragem e a função de pesos. Em imagens, podemos definir a função de amostragem $p(x, h, w)$, onde h e w são as quantidade de pixels considerados vizinhos a partir de um ponto central x nas orientações horizontal e vertical, respectivamente. Para o caso dos grafos, essa função é representada analogamente no conjunto de nós vizinhos $B(v_{ti}) = \{v_{tj} | d(v_{tj}, v_{ti}) \leq D\}$, onde $d(v_{tj}, v_{ti})$ representa a menor distância entre os nós v_{tj} e v_{ti} . Neste trabalho, Yan, Xiong e Lin adotaram o parâmetro fixo $D = 1$. Assim, a função de amostragem para grafos pode ser escrita como:

$$p(v_{ti}, v_{tj}) = v_{tj} \quad (9)$$

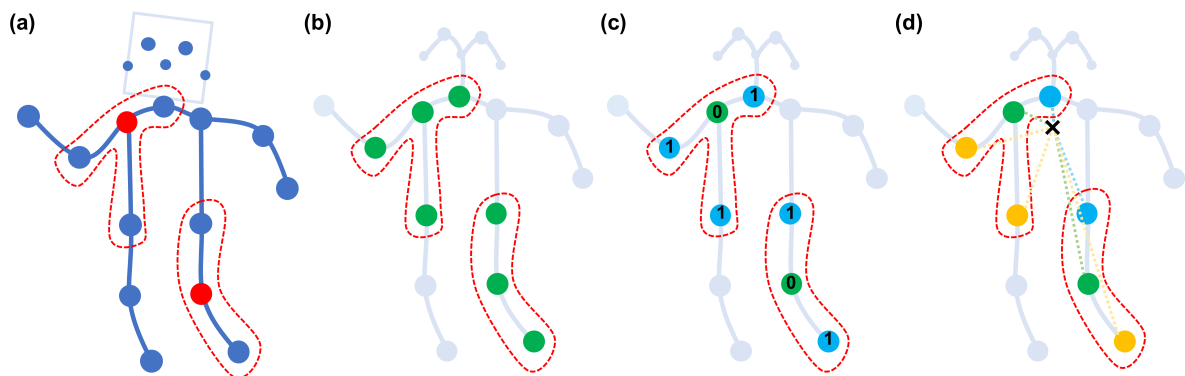
Já a função de peso, é um pouco mais complexa de ser adaptada das CNNs comuns, uma vez que nas convoluções 2D, existe uma vizinhança bem definida ao redor de determinado centro. Dessa maneira, a vizinhança dos pixels podem ser definidas em uma ordem espacial

fixa (YAN et al., 2018). Para a modelagem de grafos, a ordem é normalmente definida por um processo de rotulagem dos nós que se encontram em uma vizinhança ao redor de um nó central. No trabalho de Yan et al. (2018), essa ideia é implementada atribuindo rótulos para conjuntos $B(v_{ti})$ de nós vizinhos a um dado nó v_{ti} , particionados em um número fixo de K sub-conjuntos. Dessa forma, um rótulo l_{ti} representa por meio de um valor numérico a qual partição um dado nó pertence em relação ao nó raiz v_{ti} , tal que $l_{ti} : B(v_{ti}) \rightarrow \{0, \dots, K - 1\}$. Assim, a função de peso pode ser aplicada de acordo com a Equação 10.

$$w(v_{ti}, v_{tj}) = w'(l_{ti}(v_{tj})) \quad (10)$$

Yan et al. (2018) propõem três maneiras diferentes de particionar os nós dos grafos, todas elas são ilustradas na Figura 13. O primeiro grafo (a) representa o esqueleto de entrada para um dado quadro de um vídeo. Os nós representando as juntas estão em azul, e a vizinhança para um nó central (em vermelho) é representada pela linha pontilhada, utilizando $D = 1$. O segundo grafo (b) ilustra a primeira estratégia de partição, onde todos os nós dentro de uma vizinhança recebem o mesmo rótulo. Já no grafo (c), utiliza-se o particionamento por distância, onde os nós centrais com distância nula são rotulados em verde, e os pontos da vizinhança com distância 1 estão em azul. Por fim, o grafo (d) ilustra a estratégia de particionamento espacial, onde os nós são classificados de acordo com sua distância em relação ao centro de gravidade do esqueleto (x). Os nós raiz são representados em verde, os nós com menor distância em azul, e os com maior distância em amarelo.

Figura 13 – Estratégias de particionamento propostas para a construção das operações de convolução: (a) quadro de entrada com juntas corporais (azul) e campo receptivo (vermelho tracejado); (b) particionamento por rótulo único; (c) particionamento por distância; (d) particionamento por configuração espacial.



Fonte: Adaptada de Yan et al. (2018).

Ao performar as ações, as juntas do corpo humano se movem naturalmente de forma agrupada, e por conta disso, uma mesma junta pode aparecer em diferentes partes do corpo. No entanto, dependendo da parte do corpo analisada, a mesma junta representa um grau de importância diferente para o movimento em questão (YAN et al., 2018). A fim de modelar essa

dinâmica, o método de Yan et al. (2018) propõe o uso de uma máscara M com parâmetros treináveis em cada camada de convolução espaço-temporal do grafo, de uma forma que a contribuição de cada nó para seus nós vizinhos é escalada de acordo com o peso de importância aprendido para cada aresta espacial do grafo contida em E_S .

Resultados Experimentais

O trabalho de Yan et al. (2018) utiliza um dos maiores conjuntos de dados para a tarefa de reconhecimento de ações humanas, que contém anotações das juntas em 3 dimensões, a NTU-RGB+D (WANG et al., 2022b). Detalhes desta base de dados são apresentados na Seção 5.1

A Tabela 2 mostra resultados obtidos pelo modelo de ST-GCN (YAN et al., 2018), bem como por outros modelos estado-da-arte da época. As acurácias são obtidas a partir do reconhecimento *top-1* das ações, realizado nas duas configurações propostas pelo autor da base de dados (X-Sub e X-View). O modelo de Yan et al. (2018) apresentou os melhores resultados em ambas as configurações, além de contar com uma arquitetura mais simples que os outros modelos, e sem utilizar nenhum tipo de técnica de aumento de dados. Vale ressaltar que em todos os casos, foram utilizadas poses 3D capturadas no conjunto de dados utilizando Kinect.

Tabela 2 – Resultados obtidos pelo modelo de Yan et al. (2018).

Modelo	Acurácia (X-Sub)	Acurácia (X-View)
Lie Group (VEERIAH et al., 2015)	50.1%	52.8%
H-RNN (DU et al., 2015)	59.1%	64.0%
Deep LSTM (WANG et al., 2022b)	60.7%	67.3%
PA-LSTM (WANG et al., 2022b)	62.9%	70.3%
ST-LSTM+TS (LIU et al., 2016)	69.2%	77.7%
Temporal Conv (KIM; REITER, 2017)	74.3%	83.1%
C-CNN + MTLN (KE et al., 2017)	79.6%	84.8%
ST-GCN	81.5%	88.3%

Fonte: Adaptada de Yan et al. (2018).

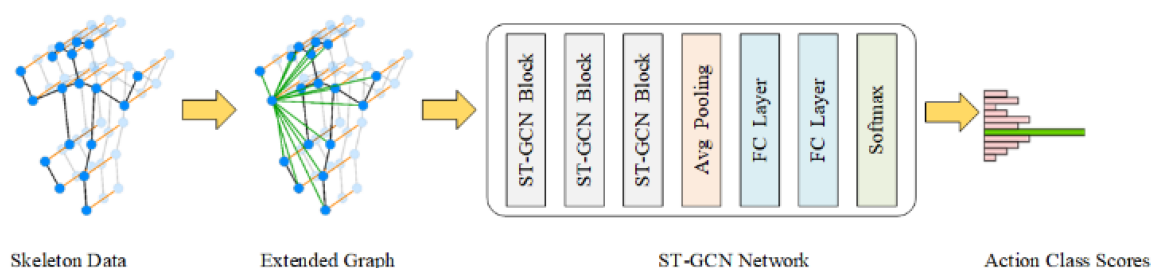
3.4.2 Extensão do Modelo Baseado em Grafo Espaço-Temporal

O trabalho de Yan et al. (2018) trouxe importantes avanços para a área de reconhecimento de ações humanas ao introduzir os grafos espaço-temporais. No entanto, a topologia convencional estática dos grafos é incapaz de mapear os relacionamentos implícitos entre nós não-adjacentes, e que podem conter informações essenciais do movimento durante a execução de uma ação. Além do problema na representação espacial, a estratégia de particionamento dos nós descartam a maioria das dependências temporais entre os nós não fisicamente conectados. A fim de contornar essas deficiências do método básico de ST-GCN, Wang et al. (2022a) propõem uma topologia de grafo de poses estendida e uma estratégia de particionamento

focada na extração de informações relacionais entre juntas não-adjacentes, o que resultou em um conjunto mais robusto e discriminativo de características.

A Figura 14 ilustra o funcionamento geral do método de Wang et al. (2022a). Os grafos de poses inicialmente são convertidos para uma estrutura de grafo estendido, que em seguida alimenta uma rede de ST-GCN, que apresenta como saída a distribuição de probabilidades para cada classe de ação. A rede de GCN espaço-temporal conta com múltiplos blocos de convolução, ligados à uma camada de *pooling* por média, seguido de duas camadas densamente conectadas, que por fim alimentam a camada de *SoftMax*, responsável por calcular a probabilidade para cada classe de ação.

Figura 14 – Visão geral do funcionamento do método de ST-GCN estendido (WANG et al., 2022a).



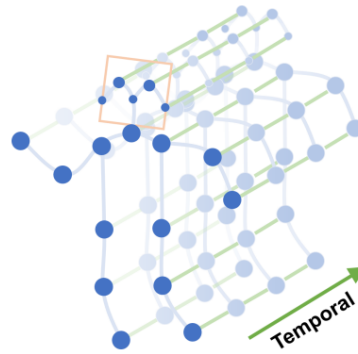
Fonte: Adaptada de Wang et al. (2022a).

Construção dos Grafos de Esqueleto Estendidos

Para obter as poses o estimador de poses OpenPose (CAO et al., 2018) foi utilizado, com um modelo de esqueleto BODY-25, que conta com 25 juntas. O método utiliza a mesma técnica apresentada na Seção 3.4.1 para a construção básica dos grafos de poses, representando os esqueletos como $G = (V, E)$, onde V é o conjunto dos vértices (nós) e E é o conjunto de arestas formado pelos ossos ligando os nós. A fim de considerar a informação temporal na representação de grafo, o método de ST-GCN original (YAN et al., 2018) adiciona as arestas temporais, que interligam os mesmos nós entre as poses detectadas em cada quadro do vídeo de entrada (Figura 15). Além disso todas as arestas do grafo são criadas com peso igual a 1, não levando em consideração as conexões implícitas entre juntas não-adjacentes (WANG et al., 2022a).

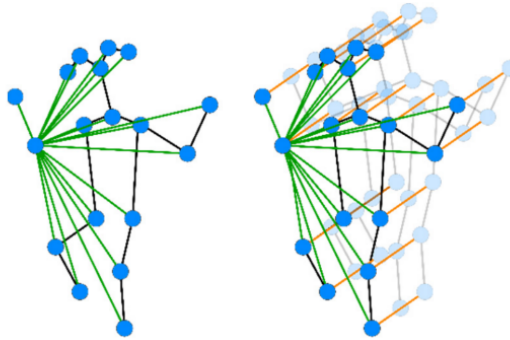
Em Wang et al. (2022a), os autores buscam representar a dinâmica amplamente conectada do movimento humano em relação a cada junta, de uma maneira em que todos os nós relacionados são conectados com arestas, além das arestas naturais (membros). Além disso, considera-se também que existem diferentes graus de relacionamento entre as juntas não conectadas diretamente quando comparadas com as juntas adjacentes. Dessa forma, os pesos das arestas são definidos de acordo com a distância das juntas, de uma forma que quanto mais distante dois vértices estão, menor será o peso das arestas entre eles. A Figura 16 ilustra a construção de um grafo estendido e sua representação espaço-temporal.

Figura 15 – Grafo espaço-temporal de uma sequência de poses proposto por Yan et al. (2018). As conexões físicas são representadas em azul, e as temporais em verde.



Fonte: Adaptada de Yan et al. (2018).

Figura 16 – Exemplo de um grafo estendido com conexões naturais em preto e conexões entre nós não adjacentes em verde (à esquerda), e a composição do grafo espaço-temporal com as conexões temporais em laranja (à direita) (WANG et al., 2022a).



Fonte: Adaptada de Wang et al. (2022a).

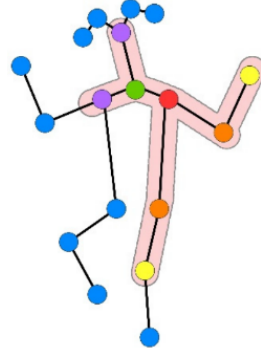
Estratégia de Partição Estendida

A maioria dos métodos para reconhecimento de ações baseados em ST-GCN utiliza a função de mapeamento apresentada na Seção 3.4.1, basicamente dividindo os nós em 3 categorias. Apesar de funcionar adequadamente, essa estratégia não considera a relação entre as juntas mais distantes ou não fisicamente conectadas (WANG et al., 2022a). Nas ST-GCNs convencionais, o método de particionamento utilizado é o espacial, mostrado na Figura 13 (d), onde existem três categorias de nós: os centrais, os concêntricos (mais próximos do centro de gravidade) e os excêntricos (mais distantes do centro de gravidade).

O esquema de particionamento proposto em Wang et al. (2022a) divide as partes do corpo entre movimentos concêntricos e excêntricos. O centro de gravidade do esqueleto é obtido pela média das coordenadas dos nós, e os nós ao redor do centro são divididos em 5 categorias, representadas pelo conjunto $\rho = \{\tau = 0, \xi = 1, \phi = 2, \lambda = 3, \mu = 4\}$, onde τ é o ponto central, ξ são os nós próximos do ponto central e mais próximos do centro de gravidade que o ponto central, ϕ são os nós próximos do ponto central mas mais distantes do centro de gravidade que o nó central, λ são os nós distantes do nó central mas mais próximos do centro

de gravidade que o nó central e por fim μ são os nós distantes do nó central e mais distantes do centro de gravidade que o nó central. A Figura 17 ilustra a estratégia de particionamento proposta em Wang et al. (2022a).

Figura 17 – Esquema de particionamento estendido: nó τ em vermelho, nós ξ em verde, nós ϕ em laranja, nós λ em roxo e nós μ em amarelo (WANG et al., 2022a).



Fonte: Adaptada de Wang et al. (2022a).

Resultados Experimentais

A avaliação do desempenho da extensão do ST-GCN proposta em Wang et al. (2022a) foi realizada utilizando o mesmo conjunto de dados NTU-RGB+D (WANG et al., 2022b) apresentado em maiores detalhes na Seção 5.1. O método foi avaliado nas duas configurações *X-Sub* e *X-View* da base de dados, utilizando a acurácia de reconhecimento *rank-1*. A Tabela 3 exibe a performance obtida pelo método de ST-GCN utilizando as extensões de estratégia de partição e de geração do grafo de poses, chamadas de rede convolucional de grafo estendido de 2 fluxos, do inglês *2-stream extended graph convolutional network* (2S-EGCN), em comparação com outros métodos de reconhecimento de ações existentes.

Tabela 3 – Resultados obtidos pelo modelo de Wang et al. (2022a).

Modelo	Acurácia (<i>X-Sub</i>)	Acurácia (<i>X-View</i>)
Deep LSTM (WANG et al., 2022b)	60.7%	67.3%
ST-GCN (YAN et al., 2018)	81.5%	88.3%
AS-GCN (LI et al., 2019)	86.8%	94.2%
2S-AGCN (SHI et al., 2019)	88.5%	95.1%
2S-EGCN	89.1%	95.5%

Fonte: Adaptada de Wang et al. (2022a).

Observando os resultados, é possível notar que as extensões propostas proporcionaram uma grande melhora na performance do método original de ST-GCN (YAN et al., 2018) em ambas as configurações de validação do conjunto de dados, apresentando uma melhora de acurácia de 9,3% e 8,2% nas configurações *X-Sub* e *X-View*, respectivamente. Além disso, o modelo supera os resultados de acurácia até então alcançados por outros métodos baseados em GCN existentes.

3.4.3 Modelo de Arquitetura Híbrida

Apesar do avanços proporcionados pelos métodos de reconhecimento de ações baseados em grafos de poses, ainda existem dificuldades intrínsecas à esse tipo de abordagem, como a incapacidade de mapear as variações nas relações de correlação entre as juntas do corpo para diferentes tipos de ação, uma vez que uma única matriz de adjacências é utilizada. Outra dificuldade é o mapeamento de relações entre juntas adjacentes entre diferentes *frames*, já que as arestas entre-quadros podem apenas conectar as mesmas juntas (YANG et al., 2023).

A fim de superar essas limitações, Yang et al. (2023) propõem uma integração entre CNNs e GCNs para o reconhecimento de ações baseado em poses, chamada de *Hybridnet*. Nesta abordagem, as CNNs e GCNs são utilizadas como complementares, aproveitando-se da capacidade das CNNs de gerarem conexões espaço-temporais densas em diferentes canais de correlação, unida à capacidade das GCNs de abstrair informações valiosas a partir da estrutura natural do esqueleto humano.

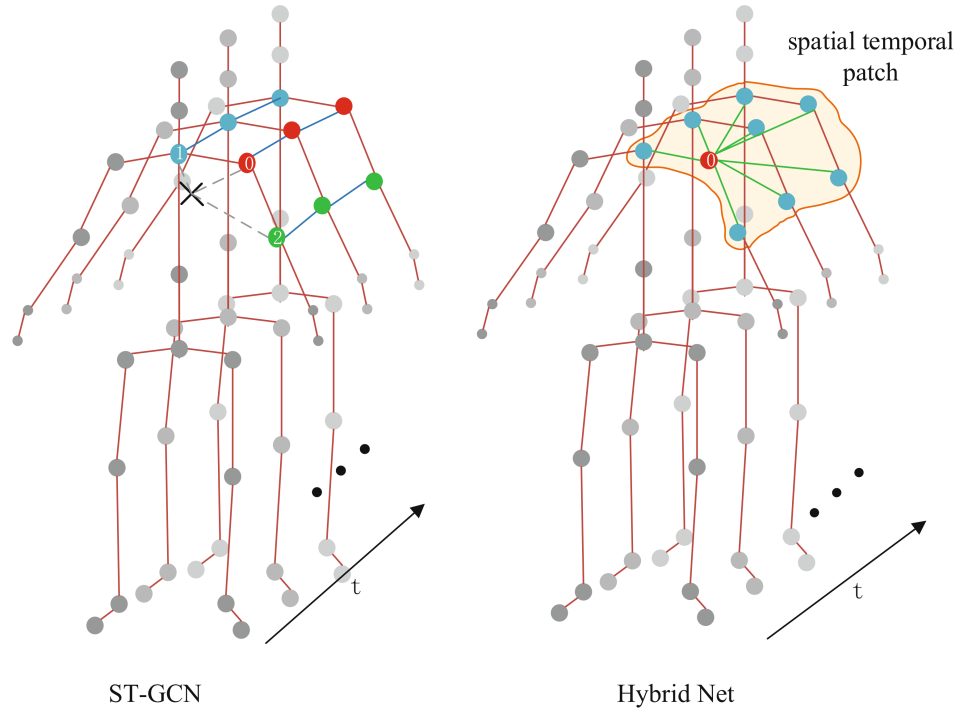
O modelo recebe como entrada uma sequência de esqueletos, que alimentam uma camada de extração de características baseado em GCN, responsável por modelar as informações estruturais, bem como a topologia dos nós do grafo. Em seguida, a informação é enviada para uma camada responsável por gerar duas componentes de características, a local e a global. A componente global proporciona informação global dos nós ao mesmo tempo que previne a perda de informação estrutural do grafo para a representação matricial 2D. Já a componente local é responsável por capturar informações locais dos nós e transmitir a informação estrutural para as próximas camadas (YANG et al., 2023).

A saída dos componentes locais e globais são concatenadas e alimentam a camada de extração de características baseada em redes neurais convolucionais. O módulo baseado em CNN é responsável pela captura de informação espaço-temporal, de uma maneira abundante, sendo capaz de reconhecer dependências complexas entre juntas em diferentes ações. As saídas do módulo de GCN e de CNN são concatenadas e passam por operações de *pooling* global, que por fim alimentam uma camada densamente conectada, seguida por uma camada de *Softmax*, responsável por obter a distribuição de probabilidades da amostra pertencer a cada classe de ação (YANG et al., 2023). Diferente do método de ST-GCN e suas extensões, nos quais apenas os mesmos nós são ligados entre diferentes *frames*, o *Hybridnet* é capaz de modelar as relações entre diferentes nós em diferentes *frames* de maneira dinâmica, como mostra a Figura 18.

Resultados Experimentais

Muitos métodos recentes de reconhecimento de ações baseado em poses utilizam uma estratégia de fusão *multi-stream*, na qual diferentes tipos de informação são utilizados, de maneira que seus *scores* são unidos por meio de uma soma ponderada a fim de obter o resultado final da classificação. No caso da *Hybridnet* (YANG et al., 2023), quatro informações foram

Figura 18 – Comparação entre os grafos espaço-temporais utilizado pelo método ST-GCN (esquerda) (YAN et al., 2018) e *Hybridnet* (direita) (YANG et al., 2023).



Fonte: Adaptada de Yang et al. (2023).

utilizadas, que são: as juntas, os ossos (ligações entre as juntas) e o movimento das juntas e dos ossos, respectivamente. Dessa forma, foram usadas três variações do método para comparação com outras abordagens: *1s-Hybridnet* que utiliza apenas informação das juntas, *2s-Hybridnet* que utiliza a informação das juntas e dos ossos, e *4s-Hybridnet* que utiliza as informações das juntas, dos ossos e seus respectivos movimentos.

Para a avaliação dos resultados, foi utilizada a base de dados NTU-RGB+D (WANG et al., 2022b), com as duas configurações propostas pelo autor, *X-Sub* e *X-View* apresentadas na Seção 5.1. A métrica utilizada foi a acurácia *rank-1*. A Tabela 4 exibe os resultados obtidos pela *Hybridnet* (YANG et al., 2023) em comparação com outros métodos estado-da-arte para reconhecimento de ações baseados em GCN.

Tabela 4 – Resultados obtidos pelos modelos *Hybridnet* de Yang et al. (2023).

Modelo	Acurácia (<i>X-Sub</i>)	Acurácia (<i>X-View</i>)
Deep LSTM (WANG et al., 2022b)	60.7%	67.3%
ST-GCN (YAN et al., 2018)	81.5%	88.3%
AS-GCN (LI et al., 2019)	86.8%	94.2%
2S-AGCN (SHI et al., 2019)	88.5%	95.1%
1s-Hybridnet	88.3%	95.7%
2s-Hybridnet	90.4%	96.5%
4s-Hybridnet	91.4%	96.9%

Fonte: Adaptada de Yang et al. (2023).

Observando a Tabela 4, é possível verificar que o método de Yang et al. (2023) obteve os melhores resultados de acurácia, principalmente utilizando a técnica de fusão de *multi-streams*. Os resultados mostraram que ao unir as arquiteturas de GCN e CNN para a tarefa de reconhecimento de ações, suas vantagens e desvantagens se complementam, obtendo resultados superiores às técnicas puramente baseadas em CNNs.

3.4.4 Abordagens Baseadas em *Transformers*

O grande sucesso do uso de *transformers* em tarefas que envolvem o processamento sequencial de dados, como modelos baseados em linguagem natural, fomentou o desenvolvimento de modelos que usassem *transformers* para tarefas de reconhecimento de imagens, que por muitas vezes superavam o desempenho das redes neurais convolucionais (CNN) (ULHAQ et al., 2022). Nesse contexto, os *transformers* aplicados à visão computacional se tornaram uma solução promissora e atrativa para problemas complexos de reconhecimento em vídeos, como o reconhecimento de ações humanas.

Uma das abordagens mais clássicas utilizadas para processamento sequencial de dados, como o Processamento de Linguagem Natural (PLN), são as Redes Neurais Recorrentes, do inglês *Recurrent Neural Networks* (RNN) (ULHAQ et al., 2022). Apesar do sucesso dessas abordagens, a arquitetura possui um problema relacionado ao esquecimento de informações recentes no processamento de sequências relativamente grandes.

O problema de esquecimento das RNNs motivou os pesquisadores da área a encontrarem uma solução, que levou ao desenvolvimento de um mecanismo que fosse capaz de dar maior "atenção" à estados importantes da sequência, evitando que os mesmos sejam esquecidos no decorrer de uma sequência longa. Esse mecanismo é conhecido como Atenção, e por muitas vezes considerado sinônimo de *Transformers* (ULHAQ et al., 2022).

Transformer é um modelo de aprendizado profundo proposto inicialmente por Vaswani et al. (2017), apresentando uma abordagem relativamente nova em comparação com as técnicas mais clássicas. Esse modelo tem sido amplamente aplicado em problemas de processamento de linguagem natural, visão computacional e processamento de fala. A primeira arquitetura foi inicialmente proposta para tratar problemas de tradução de sequência-para-sequência, no entanto, com o desenvolvimento de pesquisas na área, diversas variações do *Transformer* foram propostos posteriormente (LIN et al., 2021). Essas modelagens alternativas buscam aplicar melhorias à abordagem original, chamada atualmente de *Vanilla Transformer*, em diferentes pontos como:

- **Eficiência.** O mecanismo de auto-atenção utilizado no *Transformer* possui uma alta complexidade em termos de computação e memória, o que pode ser um empecilho. As melhorias propostas envolvem variações como mecanismos de atenção esparsa, recorrência e hierarquia;

- **Generalização.** *Transformers* são modelos amplamente flexíveis, o que faz com que seja difícil treinar em conjuntos de dados menores. Neste contexto, diversas técnicas como regularização e pré-treinamento vêm sendo propostas.

Atenção é o conceito básico sob o qual as arquiteturas de *Transformers* são construídas. Uma função de atenção pode ser descrita como mapeamento de uma consulta e um conjunto de pares chave-valor para uma saída, no qual a consulta, chaves, valores e saídas são todos vetores. A saída é calculada como uma soma ponderada dos valores, na qual os pesos são obtidos por uma função de compatibilidade entre a consulta e as chaves correspondentes (VASWANI et al., 2017). O conceito de atenção surgiu inicialmente no contexto das RNNs (ULHAQ et al., 2022).

De maneira geral, o mecanismo de atenção proposto para RNNs computa um vetor de contexto c_j , que é utilizado em conjunto com o estado anterior para definir o próximo estado da sequência. De acordo com Ulhaq et al. (2022), o vetor de contexto é obtido a partir do cálculo das similaridades entre os estados do *encoder* $h_i, \forall i \in \{1, \dots, m\}$, onde m é a quantidade de pontos de dados contidos na sequência, e o *decoder* s_j . Geralmente, para obter a similaridade, calcula-se primeiro os valores das chaves $k_i = W_K h_i$ e as consultas $q_j = W_Q s_j$, onde W_K e W_Q são matrizes de parâmetros a serem aprendidos. Dessa forma, a similaridade entre uma chave e uma consulta pode ser obtida calculando $a_i^j = k_i^T q_j$. Por fim, o vetor de contexto pode ser calculado de acordo com a Equação 11,

$$c_j = \sum_{i=1}^m a_i^j v_i \quad (11)$$

onde $v_i = W_V h_i$, e v_i é o vetor de valores, e W_V uma matriz de parâmetros.

Neste contexto, Vaswani et al. (2017) propôs uma maneira de estimar os parâmetros W_K , W_Q e W_V a partir da sequência de entrada, de uma maneira que o modelo utilize informações da sequência completa e de maneira simultânea para os vetores de contexto, consequentemente removendo a recorrência na modelagem da sequência como era feito nas redes neurais recorrentes, e dando origem às arquiteturas conhecida hoje como *Transformers* (ULHAQ et al., 2022).

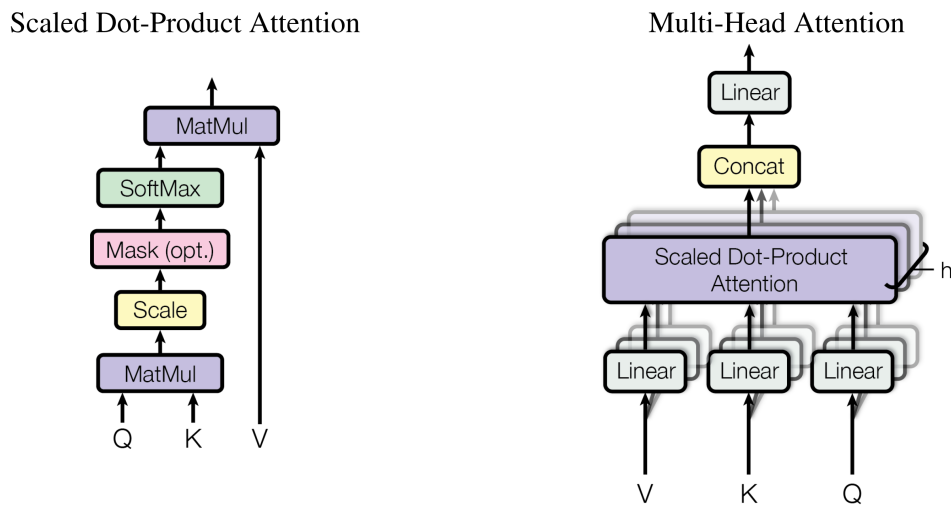
O mecanismo de atenção utilizado no *Vanilla Transform* é chamado de *Scaled Dot-Product Attention* (Figura 19). A entrada deste módulo consiste em consultas e chaves de dimensão d_k e valores de dimensão d_v . São calculados os produtos escalares da consulta com todas as chaves, dividindo cada um por $\sqrt{d_k}$, e aplicando uma função de *softmax* para obter os pesos dos valores. Na prática, a função de atenção é aplicada para um conjunto de consultas de maneira simultânea, representadas em uma matriz Q , e as chaves e valores representadas pelas matrizes K e V , respectivamente (VASWANI et al., 2017). Dessa forma, a matriz de

saídas pode ser obtida de acordo com a Equação 12. Assim, com a ajuda do vetor de contexto c_j , todos os estados do *encoder* influenciam nos estados do *decoder*.

$$Attention(Q, K, V) = V \cdot softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (12)$$

Ao invés de aplicar uma única função de atenção com as chaves, valores e consultas de dimensão d_{model} , Q, V e K são projetados linearmente h vezes, com diferentes projeções lineares para as dimensões d_k e d_v . Para cada versão projetada das consultas, chaves e valores, a função de atenção é aplicada em paralelo, produzindo uma saída de valores de dimensão d_v , que por sua vez são concatenados e projetados novamente, resultando nos valores finais (Figura 19). Esse mecanismo de *Multi-Head Attention* (atenção de múltiplas cabeças) permite que o modelo mapeie informações de diferentes subespaços de representações em diferentes posições (VASWANI et al., 2017).

Figura 19 – (Esquerda) Atenção por Produto Escalar Escalonado. (Direita) Atenção Multi-Cabeça composta por várias camadas de atenção em paralelo.



Fonte: Vaswani et al. (2017).

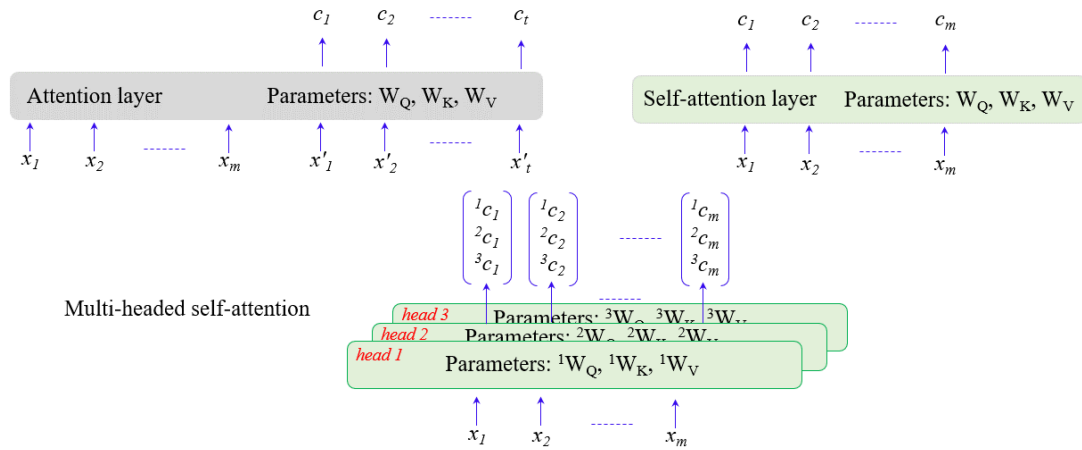
A utilização do mecanismo de *Multi-Head Attention* oferece diversas vantagens como uma representação mais rica e abrangente dos dados, uma melhor capacidade de modelagem de dependências de longo alcance, paralelização eficiente e a capacidade de aprender atenções ponderadas para diferentes partes da sequência (VASWANI et al., 2017), uma vez que os parâmetros são inicializados de maneira única em cada cabeça e são gerados vetores de contexto distintos.

3.4.5 Transformers na Visão Computacional

Um marco para o início do uso de *transformers* em tarefas de visão computacional foi o trabalho de Dosovitskiy et al. (2020), no qual os autores apresentam uma adaptação do *encoder*

proposto por Vaswani et al. (2017) para a tarefa de classificação visual, que foi chamado de *Vision Transformer* (ViT). A fim de facilitar o entendimento a respeito da adaptação proposta por Dosovitskiy et al. (2020), Ulhaq et al. (2022) apresenta uma abstração dos mecanismos de atenção em um formato de camadas de uma rede, como mostrado na Figura 20.

Figura 20 – A camada de atenção gera vetores de contexto para a modelagem sequência a sequência. Na autoatenção, múltiplas cabeças são criadas com diferentes inicializações e suas saídas são concatenadas na atenção multicabeça.



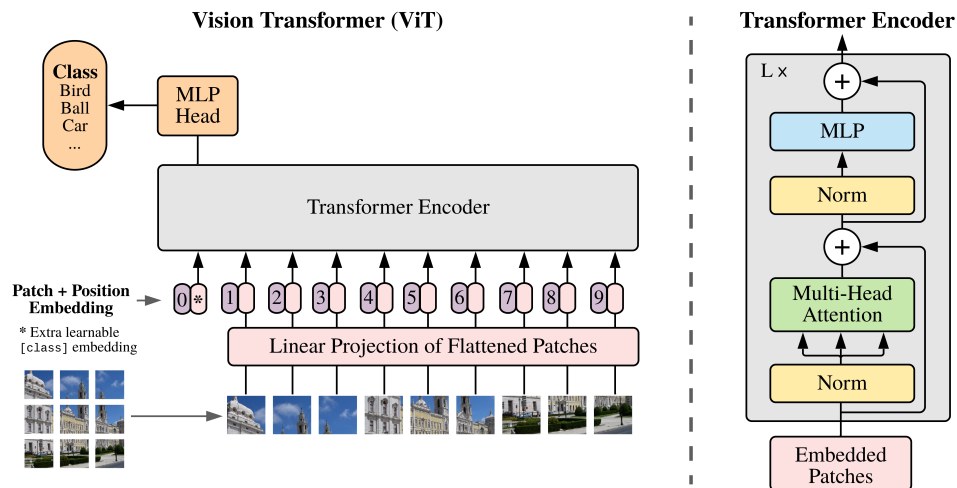
Fonte: Ulhaq et al. (2022).

De acordo com Ulhaq et al. (2022), o modelo simples de atenção é referente à modelagem de problemas de sequência para sequência baseado em *encoder* e *decoder*, calculando os vetores de contexto c_j para uma sequência $\{x'_j\}_{j=1}^t$, utilizando a sequência de entrada $\{x_i\}_{i=1}^m$. Por conta disso, a atenção simples não é muito utilizada para o contexto das tarefas de visão. Por outro lado, a camada de auto-atenção é de grande relevância para os ViT, uma vez que ao invés de relacionar uma sequência de entrada com uma sequência de saída, a auto-atenção se concentra em uma única sequência, permitindo que cada elemento "preste atenção" em outros elementos da mesma sequência. Ainda nesse contexto, o mecanismo de *Multi-Head* pode ser empregado, a fim de melhorar a representação da sequência de entrada.

A Figura 21 ilustra a arquitetura de *encoder* do ViT, proposta por Dosovitskiy et al. (2020). Esse modelo de *encoder* é diretamente inspirado no modelo de *transformer* de Vaswani et al. (2017), utilizando um mecanismo de atenção de múltiplas cabeças (*Multi-Head*). Na figura ainda é possível observar como o *Transformer Encoder* é utilizado para tarefa de classificação visual. Uma imagem de entrada é quebrada em partes menores, organizadas de forma sequencial em um vetor, que é projetado em um sub-espço numérico, muito semelhante ao processo de representação distribuída, ou *embedding* no Processamento de Linguagem Natural (PLN), onde cada elemento do vetor resultante corresponde a um valor numérico, que mapeia alguma característica específica da palavra, frase ou pedaço da imagem (no caso dos ViT (ULHAQ et al., 2022)). Além disso, uma representação posicional é atribuída a cada vetor projetado no sub-espço numérico, a fim de incorporar uma consistência espacial às

representações aprendidas. Após passar pelo *encoder*, os vetores de contexto gerados a partir das entradas são utilizados em um classificador baseado no modelo *Perceptron* de multi-camadas, do inglês *multi-layer Perceptron* (MLP).

Figura 21 – Visão geral do modelo de Dosovitskiy et al. (2020): a imagem é dividida em *patches* de tamanho fixo, linearmente embutidos com *embeddings* posicionais e processados por um *Transformer encoder*. Para classificação, utiliza-se um *token* adicional aprendível.



Fonte: Dosovitskiy et al. (2020).

No geral, as arquiteturas de *transformers* aplicadas à visão computacional realizam o cálculo da atenção, ou vetores de contexto, seguindo a Equação 12, ou variações simples da mesma. Para obter-se o valor da atenção, utilizam-se das chaves (K), consultas (Q) e valores (V), que por sua vez são calculados utilizando os parâmetros W_K , W_Q e W_V , que podem ser aprendidos de variadas maneiras nas tarefas de visão computacional, de acordo com cada modelagem proposta (ULHAQ et al., 2022).

Os ótimos resultados obtidos por Dosovitskiy et al. (2020) em tarefas de reconhecimento fomentou na área de visão computacional o uso de *transformers*, se tornando cada vez mais comum para as mais variadas tarefas como reconhecimento, detecção e geração (ULHAQ et al., 2022). Desde então, os *transformers* vêm substituindo e aprimorando as clássicas CNNs.

Outro avanço importante no uso de *transformers* para visão computacional foi a aplicação dos mesmos em tarefas envolvendo vídeos. Os vídeos podem ser definidos como uma sequência de imagens encadeadas, o que por definição já apresenta uma afinidade com o conceito básico dos *transformers*. No últimos anos, diversas abordagens foram propostas nesse contexto, em sua maioria, tomando o ViT de Dosovitskiy et al. (2020) como base, propondo novas maneiras de representar os vetores de *embeddings*, bem como a codificação posicional (ULHAQ et al., 2022). Como por exemplo Bertasius et al. (2021), que utiliza uma técnica de amostragem aleatória dos *frames* dos vídeos para serem utilizados no processo de *embedding*. Por outro lado, Arnab et al. (2021) propõem uma maneira diferente de gerar os vetores de

embedding, de uma forma que possibilita a preservação de informação temporal, baseando-se em volumes tridimensionais extraídos dos vídeos.

3.4.6 *Transformers* para Reconhecimento de Ações Humanas

A tarefa de reconhecimento de ações humanas em vídeos, se dá a partir de uma análise espacial, baseada na posição do indivíduo na cena em cada quadro do vídeo em questão e principalmente temporal, uma vez que a noção de ordem entre cada posição detectada pela câmera implica diretamente no contexto da ação em que o indivíduo executa. Duas ações opostas, como se sentar e levantar-se de uma cadeira poderiam ser interpretadas como iguais, caso a temporalidade não fosse considerada.

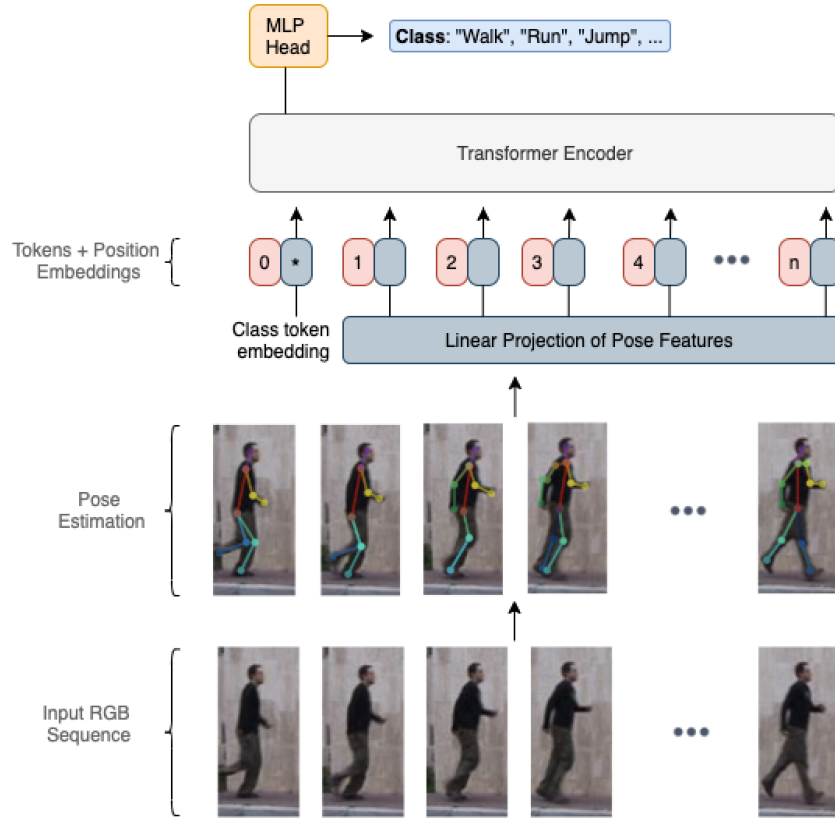
Este carácter sequencial intrínseco às ações, caracteriza uma sinergia com o conceito de *transformers*, que se mostram atualmente como uma ferramenta poderosa nesse tipo de tarefa, superando até mesmo os métodos até então considerados estado-da-arte baseados em convoluções. O reconhecimento de ações humanas pode ser abordado das mais diversas formas, tendo como principal diferencial a maneira como as ações são representadas nas imagens. Nesta dissertação, são apresentadas as abordagens baseadas em poses, que utilizam uma representação do esqueleto humano, capturando a posição de cada junta do corpo humano a cada *frame* do vídeo.

Modelo de Auto-Atenção Baseado em Poses para Reconhecimento de Ações Curtas

Inicialmente, na área de reconhecimento de ações, as abordagens baseadas em mecanismos de atenção eram construídas a partir de camadas convolucionais ou recorrentes, muitas vezes melhorando o poder de generalização desses métodos. No entanto, Mazzia et al. (2022) demonstraram como uma abordagem integralmente baseada em auto-atenção poderia superar, com uma estrutura relativamente simples, redes mistas mais complexas que ainda utilizavam convoluções ou recorrências. O método proposto, chamado de *Action Transformer*, busca reduzir ao máximo o custo computacional a partir de uma análise das poses 2D extraídas de sequências de quadros em curtas janelas de tempo, oferecendo um desempenho com capacidade para execução em tempo-real (MAZZIA et al., 2022).

A Figura 22 apresenta a arquitetura básica do método *Action Transformer* (MAZZIA et al., 2022). O *transformer* utilizado é diretamente inspirado no modelo de ViT (DOSOVITSKIY et al., 2020). O método recebe como entrada a sequência de imagens RGB, representando cada quadro de um vídeo contendo um indivíduo executando uma determinada ação. Para cada quadro, estima-se a pose do indivíduo na cena. As poses, por sua vez, são linearmente projetadas para a dimensão do modelo, que em seguida passam por uma técnica de *embedding*, gerando os *tokens* de entrada, que, por sua vez, são combinados a um *embedding* posicional, formando a entrada fornecida à camada de *encoder*. Por fim, os *tokens* de saída passam por um classificador MLP.

Figura 22 – Visão geral da arquitetura proposta por Mazzia et al. (2022): estimativas de pose são projetadas linearmente, combinadas ao *class token* e processadas pelo *Transformer encoder* com *embeddings* posicionais aprendíveis. Apenas o *class token* de saída passa por uma MLP para gerar a predição final da ação.



Fonte: Mazzia et al. (2022).

Além de propor o modelo de *Action Transformer*, Mazzia et al. (2022) apresentam um conjunto de dados chamado MPOSE2021 (MAZZIA et al., 2022), desenvolvido para o reconhecimento de ações humanas baseado em poses. O *dataset* é composto por imagens RGB de diferentes bases populares para reconhecimento de ações, como *Weizemann* (BLANK et al., 2005), *i3DPost* (GKALELIS et al., 2009), *IXMAS* (WEINLAND et al., 2006), *KTH* (SCHULDT et al., 2004), *UTKinetic-Action3D* (XIA et al., 2012), *UTD-MHAD* (CHEN et al., 2015), *ISLD* e *ISLD-Additional-Sequences* (ANGELINI et al., 2020) (MAZZIA et al., 2022). As ações contidas em cada uma das bases foram remapeadas em uma lista de 20 classes, excluindo ações que não pudessem ser mapeadas de acordo com esses 20 rótulos. Os vídeos foram divididos em 30 quadros sempre que possível, mantendo um mínimo de 20 quadros por amostra. Por fim, amostras consideradas ambíguas ou menores que 20 quadros foram descartadas, resultando em um total de 15429 amostras, em que cada uma contém um único indivíduo executando uma ação específica.

Um vídeo de entrada pode ser representado como uma sequência X_{rgb} de T quadros, de dimensão $H \times W$, representando a largura e altura dos quadros, respectivamente, com C canais de cores, tal que $X_{rgb} \in \mathbb{R}^{T \times H \times W \times C}$. Na abordagem proposta em Mazzia et al. (2022),

essa sequência passa por uma fase de pré-processamento, onde uma rede de estimação de poses 2D extrai uma sequência de poses X_{2Dpose} de dimensão $N \times T \times P$, onde N é a quantidade de pessoas presentes na cena do vídeo e P é a quantidade de pontos-chave contidos no modelo de esqueleto utilizado no estimador de poses, tal que:

$$X_{2Dpose} = F_{2Dpose}(X_{rgb}) \quad (13)$$

A camada do *Transformer Encoder* (Figura 22) recebe como entrada uma sequência unidimensional de *tokens*, de uma forma que cada uma das N sequências de matrizes $X_{2Dpose} \in \mathbb{R}^{T \times P}$ é processada de forma independente pela rede. Dessa maneira, as poses de cada um dos N indivíduos detectados na cena são processadas de forma paralela. Inicialmente, as T poses são mapeadas para um espaço de maior dimensão D_{model} utilizando uma projeção linear $W^{lo} \in \mathbb{R}^{P \times D_{model}}$. Em seguida, um vetor treinável de dimensão D_{model} é adicionado ao final da sequência de entrada, chamado de *token* de classe (CLS), que possui a função de forçar o mecanismo de auto-atenção a agregar informação em um espaço de maior dimensão, proporcionando uma representação que separa as diferentes classes de ação (MAZZIA et al., 2022). Por fim, a informação posicional é atribuída à sequência de entrada por meio de uma matriz treinável de *embedding* posicional $X_{pos} \in \mathbb{R}^{(T+1) \times D_{model}}$.

A sequência dos *tokens* linearmente projetados, juntamente com o CLS e a informação posicional é então dada como entrada para o *encoder* de L camadas, resultando na sequência de saída $X^L \in \mathbb{R}^{(T+1) \times D_{model}}$, que é a representação geral produzida pelo *transformer* em sua última camada. Na última etapa do processo, o *token* de classe CLS é isolado, e utilizado como entrada para o classificador linear baseado no *Perceptron* multi-camadas MLP_{head} , obtendo a predição da classe final pelo modelo z , de acordo com a Equação 14.

$$z = MLP_{head}(x_{cls}^L) \quad (14)$$

Durante o treinamento, apenas o *token* de classe CLS é utilizado para calcular a saída da rede MLP, enquanto todos os outros T *tokens* da sequência são utilizados apenas na entrada do *encoder*. Dessa forma, o modelo fica agnóstico à quantidade de *frames* do vídeo de entrada no momento da inferência, mesmo que seu treinamento tenha sido realizado em uma base com uma quantidade fixa T de quadros por sequência (MAZZIA et al., 2022).

Os experimentos realizados em Mazzia et al. (2022) foram todos realizados utilizando a base de dados MPOSES2021, construída a partir de outras bases clássicas para reconhecimento de ações humanas. Devido à grande heterogeneidade e ao grande número de atores únicos resultantes na base (100 atores), os autores propuseram a divisão da base em 3 cortes, chamados de *Split1*, *Split2* e *Split3*, respectivamente. Em cada um deles, 21 atores foram aleatoriamente selecionados para o teste, e o restante da base foi utilizada para treinamento. Para a extração

das poses, dois métodos de estimação de poses 2D diferentes foram utilizados: *OpenPose* (CAO et al., 2018) e *PoseNet* (PAPANDREOU et al., 2018).

O método proposto pelos autores foi dividido em quatro variações, diferenciando entre si pelos parâmetros utilizados na arquitetura do modelo, apresentadas na Tabela 5. Para fins de comparação com resultados até então tidos como estado-da-arte, foram selecionados dois métodos baseados na combinação entre convolução, mecanismo de atenção e redes recorrentes: MLSTM-FCN (KARIM et al., 2019) e *ActionXPose* (ANGELINI et al., 2020).

Tabela 5 – Configurações do modelo *Action Transformer*.

Modelo	D_{model}	D_{mlp}	L
AcT- μ	64	256	4
AcT-S	128	256	5
AcT-M	192	256	6
AcT-L	256	512	6

Fonte: Adaptada de Mazzia et al. (2022).

Os resultados obtidos pelas quatro variações do modelo proposto e dos métodos estado-da-arte selecionados para comparação, utilizando o *OpenPose* (CAO et al., 2018) na fase de extração de poses 2D, são exibidos na Tabela 6. Esses resultados mostram o potencial do uso de arquiteturas puramente baseadas em *transformers* para o reconhecimento de ações, uma vez que todas as versões propostas superaram os outros métodos, mesmo em sua versão mais compacta (*AcT* - μ). O modelo se mostrou capaz de extrair características robustas de forma generalizada, e o aumento da quantidade de parâmetros da rede refletiu diretamente na melhoria da acurácia (MAZZIA et al., 2022).

Tabela 6 – Resultados obtidos pelo modelo *Action Transformer* (OpenPose).

Modelo	Acurácia (<i>Split1</i>)	Acurácia (<i>Split2</i>)	Acurácia (<i>Split3</i>)
ActionXPose	87.60 $\pm$ 0.98	88.42 $\pm$ 0.70	89.96 $\pm$ 1.00
MLSTM-FCN	88.62 $\pm$ 0.74	90.19 $\pm$ 0.68	89.80 $\pm$ 0.94
AcT- μ	90.86 $\pm$ 0.36	91.00 $\pm$ 0.24	89.98 $\pm$ 0.47
AcT-S	91.21 $\pm$ 0.48	91.23 $\pm$ 0.19	90.90 $\pm$ 0.87
AcT-M	91.38 $\pm$ 0.32	91.08 $\pm$ 0.48	91.01 $\pm$ 0.57
AcT-L	91.11 $\pm$ 0.32	91.46 $\pm$ 0.42	91.05 $\pm$ 0.80

Fonte: Adaptada de Mazzia et al. (2022).

Modelo para Reconhecimento de Ações Baseado em *Transformer* de Dois Fluxos

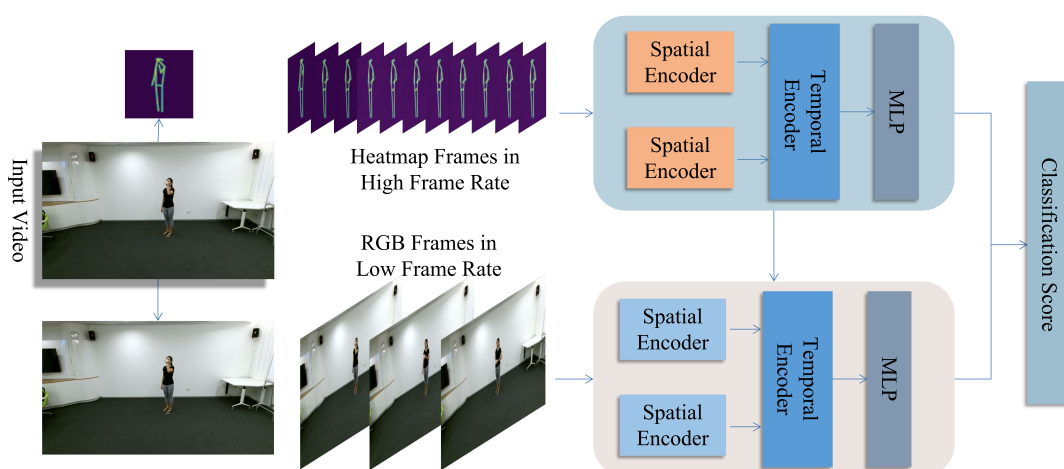
Os *transformers* têm sido amplamente utilizados para problemas de reconhecimento de ações humanas em sua grande parte baseados na modalidade de esqueleto, no qual obtém-se uma representação 2D ou 3D do corpo humano em cada quadro de um vídeo contendo uma ação. Esse tipo de abordagem tem mostrado ótimos resultados no reconhecimento de ações

simples ou com interação entre humanos, no entanto, para ações que envolvem interação entre humanos e objetos, as informações de aparência não podem ser capturadas, resultando em uma performance abaixo do ideal (SHI et al., 2023).

Neste contexto, Shi et al. (2023) propõem um algoritmo para reconhecimento de ações humanas baseado tanto nas poses estimadas a partir das imagens RGB capturadas em vídeo quanto na própria informação RGB. O algoritmo é dividido em dois fluxos, um focado na informação temporal e outro na informação espacial. De acordo com os autores, devido à falta de informação acerca da aparência, as representações de esqueleto focam nas características do movimento, o que confere um caráter mais temporal do que puramente espacial a essa abordagem. Por outro lado, as imagens RGB são usadas para mapear as informações de aparência e codificam mais informações espaciais.

A Figura 23 ilustra o funcionamento geral do algoritmo proposto por Shi et al. (2023), chamado de *RGBSformer*. Após obter os vídeos em formato RGB contendo as ações, um método de estimação de poses é utilizado para obter os esqueletos e gerar uma sequência de mapas de calor a mais alta taxa de quadros. Em seguida, os mapas de calor de esqueletos e os quadros RGB alimentam o *RGBSformer* com diferentes dimensões temporais e espaciais distintas. Após a decomposição, os dois fluxos passam por um codificador espacial, que gera as representações a nível de *frame*, que por sua vez alimentam uma camada de codificador temporal seguida por uma camada de *Perceptron* multicamadas. Por fim, após passar pela etapa de classificação, uma média é obtida a partir dos *scores* de cada fluxo, representando o resultado de classificação final.

Figura 23 – Arquitetura do RGBSformer: dois fluxos complementares, RGB (lento) para características espaciais e esqueleto (rápido) para informação temporal, fundidos em *Transformer* para a predição de ações (SHI et al., 2023).



Fonte: Shi et al. (2023).

Para avaliar o desempenho do método proposto, Shi et al. (2023) utilizaram o conjunto de dados NTU RGB+D (WANG et al., 2022b), que conta com vídeos gravados em um ambiente interno, contendo cerca de 57 mil vídeos de 60 classes de ações diferentes. Os experimentos

foram realizados utilizando a estratégia de partição inter-indivíduos (*X-Sub*) proposta pelos autores da base de dados. A Tabela 7 exibe os resultados obtidos pelo modelo proposto em comparação com outros métodos estado-da-arte baseados em GCNs e *Transformers*. Observando os resultados, é possível verificar a eficiência do uso de informações de aparência obtidos a partir do fluxo de imagens RGB unido à informação estrutural temporal das representações de mapas de calor de esqueleto, uma vez que o método *RGBSformer* apresenta uma acurácia no mínimo 0.9% maior comparado aos outros métodos.

Tabela 7 – Resultados obtidos pelo modelo de Shi et al. (2023).

Modelo	Abordagem	Acurácia
GVFE+AS-GCN (PAPADOPOULOS et al., 2019)	GCN	85.3%
Mix-Dimension (PENG et al., 2020)	GCN	89.7%
PGCN-TCA (YANG et al., 2020)	GCN	88.0%
ST-TR-AGCN (PLIZZARI et al., 2021)	<i>Transformer</i>	90.3%
EfficientGCN-B0 (SONG et al., 2022)	GCN	90.2%
RGBSformer	<i>Transformer</i>	91.1%

Fonte: Adaptada de Shi et al. (2023).

4 Abordagem Proposta

Neste capítulo é apresentada a abordagem proposta nesta dissertação para o reconhecimento de ações humanas baseada em poses 3D obtidas de poses 2D.

4.1 Visão Geral da Abordagem

Este trabalho tem como objetivo abordar os desafios do reconhecimento de ações humanas em cenários onde não há disponibilidade de informação de profundidade, utilizando apenas vídeos RGB como entrada. A partir da estimação de poses 3D com base em pontos-chave 2D, o *pipeline* proposto permite uma representação espacial mais precisa das articulações humanas, aprimorando o reconhecimento em cenários mais complexos, sem a necessidade de sensores de profundidade especializados ou outros *hardwares* complexos, tornando a solução mais prática e economicamente viável.

A Figura 24 apresenta um diagrama em blocos da abordagem proposta nesta dissertação para o reconhecimento de ações humanas utilizando poses 3D estimadas a partir de imagens RGB monoculares.

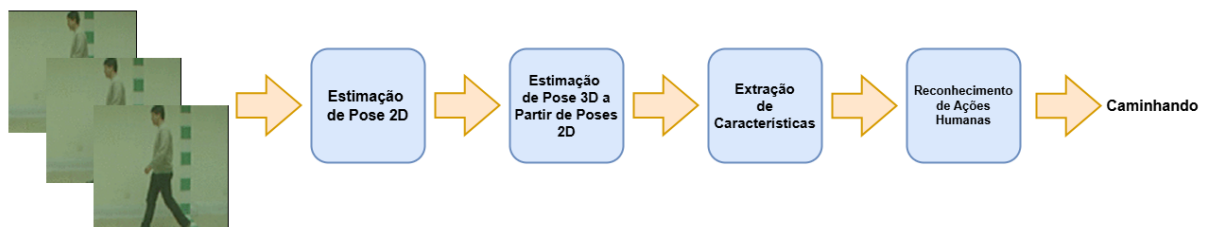


Figura 24 – Diagrama em blocos da abordagem proposta nesta dissertação para reconhecimento de ações humanas com poses 3D estimadas a partir de imagens RGB monoculares.

A abordagem proposta é composta por quatro etapas principais:

Estimação de Pose 2D: Nesta etapa, as poses 2D são estimadas a partir do vídeo RGB de entrada utilizando um método amplamente conhecido, o OpenPose, descrito na Subseção 2.2.1;

Estimação de Pose 3D: Nesta etapa, as poses 3D são estimadas a partir de imagens RGB monoculares utilizando o método DUA, descrito na Subseção 2.3.4;

Extração de Características: Nesta etapa, as características são extraídas das poses 3D, conforme descrito na Seção 4.2;

Reconhecimento de Ações Humanas: Nesta etapa, a ação humana é reconhecida por meio de um dos seguintes classificadores: 1D-CNN, ST-GCN ou ST-GCN com informações de ângulo e distância, descritos na Seção 4.3.

4.2 Extração de Características

Após a estimação de pose, são selecionados os pontos-chave mais importantes para serem usados na etapa de extração de características, ou seja, os pontos-chave que codificam mais informações da ação: 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13 e 14. Esses pontos-chave definem os braços esquerdo e direito, os antebraços, as pernas e as coxas, de acordo com o formato BODY-25 apresentado na Figura 5a.

Esta seção apresenta os métodos de extração de características das poses propostos neste trabalho.

4.2.1 Descritores Baseados em Ângulos e Distâncias das Juntas

O primeiro método de extração de característica proposto nesta dissertação é inspirado no método de reconhecimento de marcha desenvolvido por Jangua and Marana (2021).

Neste método, para cada quadro, as coordenadas 3D de todos os pontos articulares do esqueleto (pose 3D estimada) são utilizadas para calcular a angulação de cada membro em relação ao eixo horizontal e a distância entre a linha definida pelos dois pontos que representam as articulações de um determinado membro e o ponto que representa o pescoço.

Em seguida, essas duas informações (ângulos e distâncias) são concatenadas em um vetor de entrada, com 16 sequências, sendo 8 para os ângulos e 8 para as distâncias.

Após, esses vetores são submetidos a um pré-processamento, passando por uma etapa de preenchimento e truncamento a fim de definir um comprimento fixo para todas as sequências, seguida por uma etapa de suavização dos sinais, a fim de atenuar ruídos, e, finalmente, por um processo de normalização dos sinais, deixando todos em distribuição com média igual a 0 e desvio padrão igual a 1. A Subseção 4.2.1 detalha as três etapas do pré-processamento.

Considerando que uma parte do corpo de ID p no quadro i pode ser representada como um vetor $u_{i,p}$ formado por dois pontos $K_{i,p1}$ e $K_{i,p2}$, tem-se:

$$u_{i,p} = K_{i,p1} - K_{i,p2} \quad (15)$$

O ângulo φ entre duas partes do corpo, identificadas por a e b , no quadro i , pode ser obtido pela Equação 16:

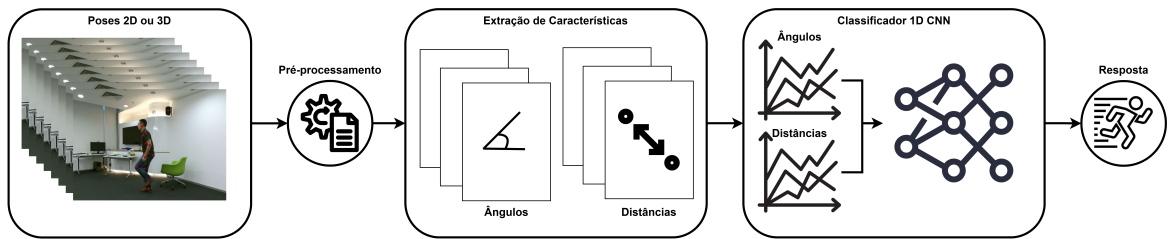
$$\varphi = \arccos \frac{u_{i,a} \cdot u_{i,b}}{\|u_{i,a}\| \cdot \|u_{i,b}\|} \quad (16)$$

De forma análoga, a distância d entre dois pontos no quadro i pode ser obtida por meio da Equação 17.

$$d = ||K_{i,p} - K_{i,1}|| \quad (17)$$

A Figura 25 mostra o diagrama de blocos do método proposto para reconhecimento de ações humanas utilizando os descritores baseados em ângulos e distâncias das juntas com o classificador 1D CNN descrito na Subseção 4.3.1.

Figura 25 – Diagrama do método proposto para reconhecimento de ações baseado em ângulos e distâncias das juntas e usando um classificador 1D CNN.



Fonte: Elaborada pelo Autor.

Etapas do Pré-Processamento

Como os vídeos podem ter quantidades diferentes de quadros, devido à variação na taxa de quadros da câmera de captura, ou no comprimento do vídeo gravado, as sequências podem ter comprimentos variados, o que é um problema, uma vez que as arquiteturas dos modelos utilizados exigem sequências de entrada de tamanho fixo. Para tratar esse problema, é realizada uma primeira etapa de pré-processamento que se baseia no preenchimento seguido de truncamento das sequências.

No primeiro passo, preenche-se as sequências repetindo o último elemento de cada uma, de maneira que todas tenham o mesmo comprimento da sequência mais longa da base de dados. Em seguida, todas as sequências resultantes são truncadas de acordo com a média dos comprimentos de todas as sequências iniciais.

A segunda etapa do pré-processamento consiste na aplicação do filtro passa-baixa de Butterworth (BUTTERWORTH, 1930), que filtra os sinais no domínio das frequências, eliminando partes em que o mesmo apresenta uma frequência maior que um determinado limiar α (neste caso $\alpha = 2Hz$), suavizando o sinal e reduzindo ruídos causados por erros de detecção na fase de estimação de pose.

Por fim, na última etapa de pré-processamento, as sequências passam por um processo de padronização, de forma que a distribuição dos valores das sequências possua média $\mu_n = 0$ e desvio padrão $\sigma_n = 1$. Para isso, deve-se, inicialmente, calcular a média (μ_i) e o desvio padrão (σ_i) do conjunto de dados de treinamento, em seguida pode-se calcular o valor x_n da sequência normalizada a partir do valor x_i da sequência inicial de acordo com a Equação 18:

$$x_n = \frac{x_i - \mu_i}{\sigma_i} \quad (18)$$

4.2.2 Descritores Baseados em Ângulos Medidos em Quatérnios

Neste método, os ângulos e distâncias são calculados de forma semelhante ao método descrito na Seção 4.2.1. No entanto, os ângulos agora são representados por quatérnios em vez de radianos, e são calculados entre cada par de membros (excluindo movimentos detalhados de mãos e pés).

Quatérnios são uma forma matemática de representar orientações e rotações 3D de forma mais estável e eficiente do que ângulos de Euler ou matrizes de rotação. Em vez de usar apenas coordenadas 3D das articulações, pode-se representar as orientações relativas de membros (como braço ou perna) por meio de quatérnios.

Aqui, um membro é representado como um vetor 3D. Dado dois vetores normalizados $\mathbf{u}_{i,a}, \mathbf{u}_{i,b} \in \mathbb{R}^3$, o quatérnio q que representa a rotação de $\mathbf{u}_{i,a}$ para $\mathbf{u}_{i,b}$ é definido como:

$$\mathbf{v} = \mathbf{u}_{i,a} \times \mathbf{u}_{i,b}, \quad w = \sqrt{\|\mathbf{u}_{i,a}\|^2 \|\mathbf{u}_{i,b}\|^2 + \mathbf{u}_{i,a} \cdot \mathbf{u}_{i,b}} \quad (19)$$

$$q = \frac{(w, \mathbf{v})}{\sqrt{w^2 + \|\mathbf{v}\|^2}} \quad (20)$$

Para o cálculo das distâncias, utiliza-se a norma Euclidiana entre cada par de pontos-chave $K_{i,p}$ e $K_{i,q}$, conforme descrito na Equação 17, onde i é o índice do quadro, e p e q são os índices dos pontos-chave. Dado um total de N_k pontos-chave, obtêm-se N_k^2 distâncias par-a-par.

4.2.3 Descritores Baseados em Grafos

Em uma GCN (*Graph Convolutional Network*), os vetores de características representam as informações iniciais ou propriedades dos nós e/ou arestas de um grafo. Esses vetores são os dados de entrada que a GCN utiliza para aprender e reconhecer padrões, como no reconhecimento de ações humanas.

Cada nó v_i em um grafo possui um vetor de características $\mathbf{x}_i \in \mathbb{R}^{d_x}$, onde:

- d_x é o número de características (ou dimensões).
- $\mathbf{x}_i$ contém valores que descrevem o nó i .

Esses vetores compõem uma matriz de características $\mathbf{X} \in \mathbb{R}^{N_v \times d_x}$, onde N_v é o número total de nós no grafo. De forma análoga, cada aresta a_i em um grafo possui um vetor de características $\mathbf{y}_i \in \mathbb{R}^{d_a}$, onde:

- d_a é o número de características (ou dimensões).
- $\mathbf{y}_i$ contém valores que descrevem o nó i .

Neste trabalho foram utilizadas duas implementações de GCN: a ST-GCN básica e a ST-GCN enriquecida com informações do movimento corporal. Conforme supracitado, os vetores de características podem ser utilizados na propagação de informações tanto dos nós, quanto das arestas. Na ST-GCN básica, foi utilizado apenas a propagação pelos nós, onde $d_v = 3$ para as poses 3D e $d_v = 2$ para as poses 2D, e x_i representa as coordenadas da junta do corpo humano 2D ou 3D representada pelo nó do grafo.

Já no caso da ST-GCN enriquecida, foi utilizada a propagação por nó e por aresta, onde $d_v = 20$ para poses 3D e $d_v = 19$ para poses 2D. O vetor x_i é representado pelas coordenadas da junta referente ao nó, concatenadas com as distâncias entre a junta atual e todas as 17 juntas consideradas no esqueleto da pose. No caso das arestas, $d_a = 60$ e y_i é representado pelo ângulo em quatérnios, entre a aresta atual e todas as outras 15 arestas do grafo.

4.3 Reconhecimento da Ação

O último passo da abordagem proposta é o reconhecimento da ação humana realizado por um classificador pré-treinado para esta finalidade. Neste trabalho, foram avaliados três classificadores: 1D-CNN, ST-GCN e ST-GCN com princípios de redes de atenção em grafos (Graph Attention Networks, GAT), descritos nas Subseções 4.3.1, 4.3.2 e 4.3.3, respectivamente.

4.3.1 1D CNN

Uma das tarefas mais fundamentais na análise de dados de séries temporais, como ocorre na análise de sequência de quadros de um vídeo, é a classificação de séries temporais (*Temporal Series Classification*, TSC) (TANG et al., 2020). Devido ao seu amplo domínio de aplicações, incluindo cuidados com a saúde, previsão de tráfego, análise do mercado financeiro, entre outros, nos últimos anos tem-se incentivado a exploração de modelos de aprendizado profundo para TSC. Nesse cenário, modelos de redes neurais convolucionais unidimensionais (1D-CNN) têm alcançado resultados promissores (TANG et al., 2020). Redes neurais convolucionais unidimensionais (1D-CNNs) aprendem automaticamente características relevantes a partir de dados brutos de séries temporais, eliminando a necessidade de extração manual de atributos. Essa capacidade as torna altamente eficazes para tarefas de classificação e regressão em séries

temporais, permitindo aprendizado de ponta a ponta diretamente a partir de dados rotulados (RIZVI, 2022).

Após a obtenção dos ângulos e distâncias, conforme descrito na Seção 4.2.1, os dados podem ser interpretados como um conjunto de séries temporais, em que cada instante de tempo é representado por um quadro do vídeo original. Por fim, o conjunto de séries temporais é utilizado como entrada para uma rede 1D-CNN, com o objetivo de reconhecer a ação realizada no vídeo.

A arquitetura da 1D-CNN utilizada neste estudo é composta por três blocos convolucionais, cada um contendo uma camada *Conv1D* com tamanhos de filtro crescentes (64, 128, 256), seguidos por normalização em lotes (*batch normalization*), *max pooling* e *dropout*. Uma camada de *global average pooling* condensa as características extraídas, que são então passadas por duas camadas densas (com 128 e 40 unidades, respectivamente), com normalização e *dropout* adicionais para melhorar a capacidade de generalização. Essa estrutura permite uma extração eficaz de características temporais a partir de dados sequenciais.

4.3.2 Rede de Convolução Espaço-Temporal em Grafos

Usar GCN (Graph Convolutional Network) para reconhecimento de ações humanas baseado em esqueletos é altamente eficaz porque esse tipo de dado tem uma estrutura natural de grafo e as GCNs são projetadas exatamente para explorar relações estruturais entre elementos interconectados. As GCNs conseguem aprender relações não-lineares entre articulações próximas e distantes, sendo, portanto, melhores do que as CNNs tradicionais, que só analisam vizinhanças locais em imagens ou vídeos.

Extensões das GCNs, tais como a ST-GCN (Spatial-Temporal GCN), adicionam conexões ao longo do tempo (por exemplo, a mesma articulação em quadros consecutivos), permitindo que a rede entenda movimentos dinâmicos, e não apenas poses estáticas.

Neste trabalho, foi adotada a ST-GCN proposta por Wang et al. (2022a) para modelar as interações entre articulações esqueléticas nos domínios espacial e temporal. A rede opera sobre as coordenadas das articulações $\mathbf{X} \in \mathbb{R}^{T \times V \times C}$, onde T denota o comprimento da sequência, V o número de articulações, e C as coordenadas espaciais. As convoluções em grafos são definidas sobre uma matriz de adjacência esquelética $\mathbf{A} \in \mathbb{R}^{V \times V}$ que codifica as conexões naturais do corpo, com pesos de importância das arestas aprendíveis $\mathbf{E}_l$ aplicados em cada camada l para modular adaptativamente as relações entre articulações.

A rede combina operações espaciais em grafos com convoluções temporais usando um tamanho de *kernel* igual a 9 para modelar padrões de movimento. Cada bloco ST-GCN emprega um design residual onde as operações espaciais transformam inicialmente as relações entre articulações em representações aprendidas, seguidas por convoluções temporais que integram informações entre quadros, permitindo o aprendizado de tarefas como reconhecimento de ações.

A modelagem utilizada segue os padrões de arquitetura da rede apresentada no Capítulo 3.4.1.

Para que o método explore adequadamente as vantagens dos dados tridimensionais, todas as poses 3D são alinhadas a um sistema de coordenadas canônico utilizando uma rotação rígida derivada de marcos anatômicos. Seja $\mathbf{p}_{\text{pelvis}} \in \mathbb{R}^3$ e $\mathbf{p}_{\text{thorax}} \in \mathbb{R}^3$ as coordenadas das articulações da pelve e do tórax no quadro inicial. O vetor do tronco $\mathbf{v}_{\text{torso}} = \mathbf{p}_{\text{thorax}} - \mathbf{p}_{\text{pelvis}}$ é projetado no plano horizontal, gerando o ângulo de rotação $\theta_y = \arctan(v_x/v_z)$. Uma matriz de rotação

$$\mathbf{R}_y(\theta) = \begin{bmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{bmatrix} \quad (21)$$

pode então ser aplicada a todas as articulações $\mathbf{p}_i$ em todos os quadros, transformando a sequência de poses como $\mathbf{p}_i^{\text{rot}} = \mathbf{p}_i \mathbf{R}_y^T(\theta)$. Isso reduz o impacto do ponto de vista na predição e elimina a variância de rotação global a partir do quadro inicial. As poses normalizadas $\mathbf{X}_{\text{rot}}$ são então propagadas pela ST-GCN para extração de características espaço-temporais.

4.3.3 Rede de Convolução Espaço-Temporal em Grafos com Princípios de Atenção

Para melhorar a acurácia no reconhecimento e avaliar a contribuição de características construídas manualmente, como ângulos e distâncias, propomos neste trabalho uma abordagem que combina as duas técnicas apresentadas anteriormente. Este método segue um *pipeline* em duas etapas, extração de características e reconhecimento, na qual a extração de características é realizada como descrito na Subseção 4.2.2 e o reconhecimento é realizado por uma ST-GCN que incorpora dados de movimento e princípios de atenção em grafos, conforme descrito a seguir.

Para incorporar dados de movimento, foi proposto o ST-GCN Motion-Based (ST-GCNMB), uma arquitetura híbrida espaço-temporal que integra princípios de redes de atenção em grafos (Graph Attention Networks, GAT).

As GATs são redes neurais para grafos que introduzem o mecanismo de atenção no processo de agregação das informações dos vizinhos, permitindo que cada nó aprenda a dar mais importância aos vizinhos mais relevantes. Diferente da GCN tradicional, que trata todos os vizinhos igualmente (ou com pesos fixos via normalização da matriz de adjacência), a GAT aprende pesos de atenção que dizem quão importante é cada vizinho para um dado nó.

O modelo opera sobre um grafo do esqueleto humano ($\mathcal{G} = (\mathcal{V}, \mathcal{E})$) utilizando dois tipos de características: (i) características dos nós, compostas por coordenadas 3D e distâncias Euclidianas par-a-par, e (ii) características das arestas, que codificam informações angulares relativas entre os membros utilizando quatérnios.

O núcleo do modelo é uma camada de convolução espacial em grafos que estende os princípios de atenção do GAT ao utilizar as características das arestas para guiar o processo de agregação. Diferentemente dos GATs tradicionais, os coeficientes de atenção são derivados diretamente das características da própria aresta. Esses coeficientes são normalizados usando uma função *softmax* aplicada à vizinhança de cada nó, resultando nos pesos de atenção finais, α_{ij} .

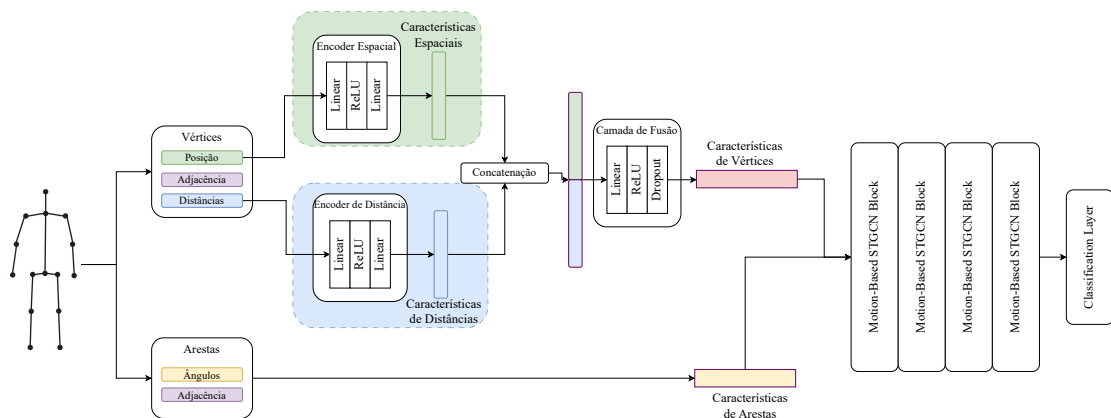
Em paralelo, as características das arestas são projetadas para o espaço de características dos nós por meio de uma transformação linear f_{prop} , gerando uma nova representação da aresta, $e'_{ij} = f_{prop}(e_{ij})$, compatível com as características dos nós. A saída final para o nó i é uma fusão entre a característica transformada do nó vizinho e essa nova representação de aresta, ponderadas pela atenção calculada:

$$y_i = \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} (W x_j + e'_{ij}) \right), \quad (22)$$

onde W é uma transformação linear aplicada às características dos nós, e σ é uma função de ativação não-linear. Essa camada espacial é então acoplada a uma convolução temporal, formando um bloco espaço-temporal completo, capaz de aprender tanto com as posições articulares quanto com os movimentos relativos do corpo.

A Figura 26 apresenta a arquitetura do modelo ST-GCNMB. Inicialmente, a estrutura esquelética é representada por vértices (articulações) e arestas (segmentos corporais). As informações dos vértices incluem posição, matriz de adjacência e distâncias euclidianas entre articulações, enquanto as arestas carregam dados sobre ângulos entre segmentos e suas conexões. Essas informações são processadas por diferentes codificadores: o *Encoder* Espacial extrai características espaciais dos vértices, e o *Encoder* de Distância captura padrões relevantes a partir das distâncias. As saídas desses *encoders* são concatenadas e passam por uma Camada de Fusão, composta por operações lineares, função de ativação ReLU e *dropout*. Paralelamente, as características das arestas são extraídas e integradas à entrada da rede principal. Ambas as representações são então processadas por blocos *Motion-Based STGCN*, que exploram correlações espaço-temporais ao longo da sequência de poses. Por fim, a saída dos blocos é encaminhada para uma Camada de Classificação, responsável por inferir a ação realizada com base nas representações extraídas.

Figura 26 – Arquitetura utilizada para o modelo de ST-GCN Motion Based (ST-GCNMB).



Fonte: Elaborada pelo Autor.

5 Resultados e Discussão

Este capítulo apresenta e discute os resultados obtidos a partir dos experimentos realizados utilizando o conjunto de dados NTU RGB+D. A avaliação foi realizada com base em métricas de acurácia, permitindo a comparação do desempenho da abordagem proposta com métodos de referência. Os resultados obtidos são analisados de forma quantitativa e qualitativa, a fim de trazer à tona discussões pertinentes a respeito do desempenho e aplicabilidade dos métodos estudados.

5.1 Base de Dados

Devido ao seu amplo uso na literatura, principalmente na validação de métodos de reconhecimento de ações considerados estado-da-arte, para este trabalho foi selecionado o NTU RGB+D, um conjunto de dados para reconhecimento de ações humanas organizado por Shahroudy et al. (2016).

O NTU RGB+D inclui dados de câmera RGB, dados de profundidade (D), dados de esqueletos humanos e mapas de infravermelho (IR), o que o torna um recurso valioso para uma ampla gama de tarefas, incluindo reconhecimento de ações, rastreamento de pose e outras aplicações de visão computacional baseadas em humanos.

Este conjunto de dados possui 56.880 amostras (vídeos) de ações realizadas por 40 indivíduos, permitindo um estudo detalhado das variações interpessoais nas ações humanas. Além disso, foram coletadas amostras de 60 diferentes categorias de ações, divididas em três grupos: ações do cotidiano (40 classes), ações relacionadas à saúde (9 classes) e ações mútuas (11 classes).

Os vídeos foram gravados utilizando três dispositivos *Microsoft Kinect V2* simultaneamente, posicionados em três ângulos: frontal, lateral e oblíquo. Isso significa que os pesquisadores têm a opção de avaliar algoritmos de estimativa de pose sob diferentes condições de visão, aumentando a robustez generalizada do algoritmo desenvolvido.

Na Figura 27, são apresentadas amostras de quadros do conjunto de dados NTU RGB+D. As quatro primeiras linhas mostram a diversidade de sujeitos humanos e vistas da câmera presentes no conjunto de dados. A quinta linha destaca a variação intraclasse das performances, exibindo como a mesma ação pode variar sutilmente quando realizada por diferentes indivíduos ou sob diversas condições. Por fim, a última linha ilustra as diferentes modalidades de dados que o conjunto NTU RGB+D oferece para um único quadro, incluindo dados de câmera RGB, câmera RGB com pontos de articulação (juntas), dados de profundidade, dados de profundidade com pontos de articulação e dados do sensor de infravermelho. Este

grande alcance de modalidades distintas reforça a versatilidade e abrangência do conjunto de dados NTU RGB+D, tornando-o um recurso valioso para o reconhecimento de ações humanas e outras tarefas relacionadas à visão computacional (SHAHROUDY et al., 2016).

Figura 27 – *Frames* extraídos do conjunto de dados NTU RGB+D. As primeiras quatro linhas mostram a variedade de indivíduos e visualizações de câmeras. A quinta linha representa a variação intraclasse das ações. A última linha ilustra as modalidades RGB, RGB+juntas, profundidade, profundidade+juntas e IR de um *frame* de amostra.



Fonte: Adaptada de Shahroudy et al. (2016).

Os criadores do conjunto de dados NTU RGB+D propuseram duas configurações de avaliação diferentes para estimar a eficácia do algoritmo de reconhecimento de ação. Nas configurações de avaliação *cross-subject* (*X-Subject*), os conjuntos de treino e teste são divididos de acordo com os indivíduos, enquanto na configuração de avaliação *cross-view* (*X-View*), o conjuntos são divididos de acordo com as diferentes câmeras, garantindo que o treinamento e teste sejam realizados em diferentes ângulos de visão.

Para os nossos experimentos, foram considerados inicialmente apenas as 40 ações

do grupo de atividades do cotidiano do NTU RGB+D, uma vez que a abordagem proposta nesta dissertação destina-se a contextos onde as ações são realizadas por apenas um indivíduo (ações que envolvem interações do indivíduo com outros indivíduos ou com objetos estão fora do escopo deste estudo). Além disso, como a abordagem proposta é baseada em poses 2D obtidas do método OpenPose, que não incluem informações detalhadas das mãos, foi necessário eliminar deste conjunto inicial de ações do cotidiano aquelas que dependem de informações minuciosas das mãos para serem corretamente classificadas. Com isso, o conjunto de ações avaliadas neste estudo ficou restrito a 10, a saber: 6 (pegar algo do chão), 7 (arremessar), 10 (bater palmas), 23 (acenar com a mão), 24 (chutar algo), 25 (colocar a mão no bolso), 26 (pular com um pé), 27 (pular para cima), 34 (esfregar as duas mãos) e 40 (cruzar as mãos à frente, sinal de pare). A Figura 28 apresenta amostras das 10 ações avaliadas neste trabalho.

5.2 Cenários dos Experimentos

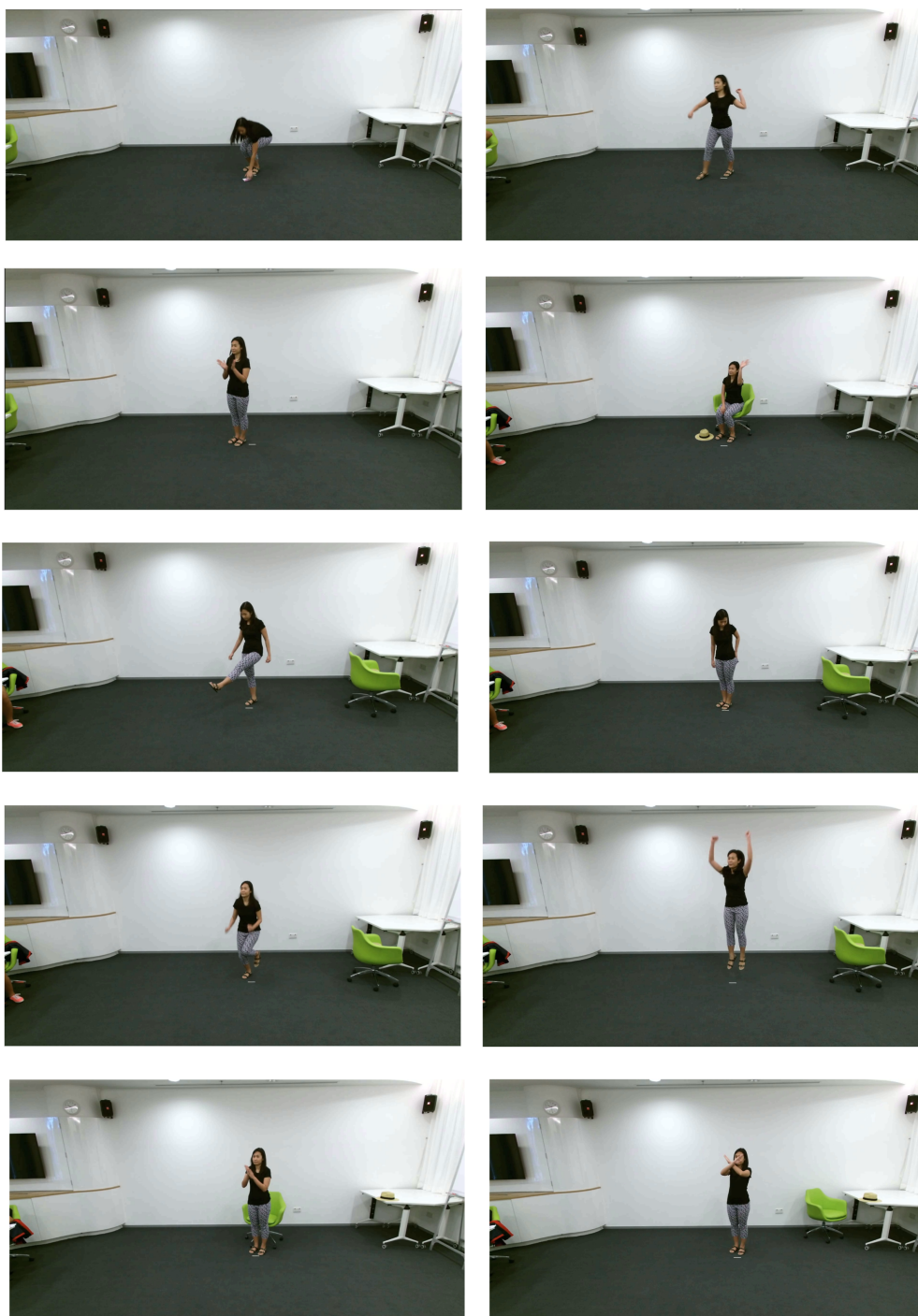
Dois protocolos experimentais padronizados foram adotados para a avaliação dos modelos: *Cross-View (X-View)* e *Cross-Subject (X-Sub)*, conforme especificado no artigo original do *NTU RGB+D Dataset* (WANG et al., 2022b). Esses protocolos foram propostos pelos autores do conjunto de dados com o objetivo de garantir uma avaliação estatisticamente consistente e comparável entre diferentes métodos da literatura. O protocolo *Cross-View* enfatiza a capacidade do modelo em generalizar para novas perspectivas de câmera, sendo o treinamento realizado com amostras capturadas pelas câmeras 2 e 3, e os testes realizados com as amostras da câmera 1, que apresenta um ângulo de visão distinto ($\pm 45^\circ$). Já o protocolo *Cross-Subject* avalia a robustez do modelo em relação às variações entre indivíduos, separando os sujeitos em dois grupos não sobrepostos — 20 para treinamento e 20 para teste — conforme a divisão definida pelos autores da base.

Ambos os protocolos foram concebidos para assegurar a reprodutibilidade experimental e permitir uma análise estatística padronizada do desempenho dos métodos. Ademais, conforme descrito no artigo original, as classes de ação na base NTU RGB+D são balanceadas, ou seja, cada classe possui um número semelhante de amostras distribuídas entre os diferentes sujeitos e pontos de vista. Esse balanceamento reduz a influência de vieses de classe e assegura que as comparações de desempenho entre modelos reflitam diferenças reais de aprendizado e generalização, e não desproporções nos dados.

5.3 Métrica de Avaliação

A métrica escolhida para avaliar o desempenho dos modelos foi a acurácia, conforme o protocolo de avaliação estabelecido pelos autores da base NTU RGB+D (WANG et al., 2022b). Essa escolha segue o padrão utilizado na literatura de reconhecimento de ações humanas baseadas em poses 3D, especialmente para esta base, na qual as classes são balanceadas e

Figura 28 – *Frames* extraídos do conjunto de dados NTU RGB+D que ilustram as 10 ações avaliadas neste trabalho. Da esquerda para a direita, de cima para baixo: (a) pegar algo do chão; (b) arremessar; (c) bater palmas; (d) acenar; (e) chutar; (f) colocar a mão no bolso; (g) pular com um pé; (h) pular para cima; (i) esfregar as duas mãos; (j) fazer o sinal de pare.



Fonte: Elaborada pelo Autor.

os experimentos seguem protocolos fixos de *Cross-View* e *Cross-Subject*. Nessas condições, a acurácia fornece uma medida direta e estatisticamente válida da proporção de previsões

corretas, sem a necessidade de ponderação adicional por classe.

Embora existam métricas mais complexas, como *precision*, *recall* ou *F1-score*, o uso da acurácia é justificado pelo fato de ela permitir comparações consistentes com os resultados reportados no artigo original e em trabalhos subsequentes que utilizam os mesmos protocolos. Além disso, a base NTU RGB+D apresenta uma distribuição equilibrada entre classes e sujeitos, o que torna a acurácia uma métrica suficiente e representativa para mensurar o desempenho global do modelo sem introduzir distorções decorrentes de desbalanceamento de dados. Assim, a adoção dessa métrica assegura compatibilidade com a metodologia de referência e possibilita uma análise quantitativa coerente com os estudos previamente publicados.

5.4 Resultados Obtidos

A Tabela 8 apresenta as acurácias obtidas pelos métodos propostos neste trabalho para descrever e reconhecer as ações humanas usando poses 3D (1D-CNN, ST-GCN e ST-GCNMB), no subconjunto composto por 10 ações do conjunto de dados NTU RGB+D, utilizando os tipos de dados disponibilizados no conjunto de dados (projeção 2D RGB, projeção 2D Depth, 3D Kinect), além das poses 3D estimadas pelo método DUA a partir de poses 2D estimadas pelo método OpenPose.

Tabela 8 – Acurácia dos métodos avaliados utilizando o conjunto de 10 ações.

Modelo	Tipo do Dado	<i>X-Sub</i>	<i>X-View</i>
1D-CNN	Projeção 2D RGB	62.06%	52.65%
1D-CNN	Projeção 2D Depth	64.61%	55.89%
1D-CNN	3D Kinect	68.48%	64.43%
1D-CNN	Poses 3D Estimadas	63.15%	58.05%
ST-GCN	Projeção 2D RGB	79.39%	78.81%
ST-GCN	Projeção 2D Depth	78.30%	79.85%
ST-GCN	3D Kinect	82.06%	82.57%
ST-GCN	Poses 3D Estimadas	80.97%	81.42%
ST-GCNMB	Poses 3D Estimadas	80.73%	84.74%

Observa-se que os modelos ST-GCN e ST-GCNMB superaram de forma consistente o modelo 1D-CNN em todos os cenários (*X-Sub* e *X-View*) e tipos de dados avaliados, sendo que todos os métodos propostos apresentaram, em geral, maiores acurácias ao utilizarem informações de pose 3D em comparação com as projeções 2D. Especificamente, o ST-GCN utilizando poses 3D estimadas obteve acurácias de 80,97% no cenário *X-Subject* e 81,42% no cenário *X-View*, superando os resultados com projeções 2D RGB, que foram de 79,39% e 78,81%, respectivamente. Tendência semelhante foi observada para o modelo 1D-CNN, que atingiu 63,15% e 58,05% com poses 3D estimadas, frente a 62,06% e 52,65% com projeções 2D RGB.

O modelo ST-GCN aprimorado, ST-GCNMB, que incorpora dados de movimento e princípios de atenção e baseados em distâncias extraídas das poses 3D estimadas, apresentou o melhor desempenho geral, alcançando 84,34% no cenário *X-View* e 80,73% no cenário *X-Subject*. Esses resultados evidenciam a eficácia do enriquecimento das representações espaciais com informações geométricas, refletindo diretamente em um ganho consistente na acurácia da classificação.

Além disso, os resultados evidenciam a vantagem da utilização de informações 3D na tarefa de reconhecimento, uma vez que o ST-GCN com poses 3D estimadas também superou o ST-GCN com projeção de profundidade em 2D. Ainda que a eficácia do ST-GCN tenha ficado ligeiramente superior com dados 3D obtidos por Kinect, a abordagem baseada em estimação reduz o custo com sensores especiais para capturar profundidade, além das demais limitações inerentes ao sensor Kinect, como a limitação de alcance.

A Tabela 9 mostra os identificadores numéricos atribuídos às classes das 10 ações do conjunto de dados utilizado e as Figuras 29 e 30 apresentam as matrizes de confusão obtidas pelos métodos baseados em GCN para o subconjunto de 10 ações, nos protocolos *X-Sub* e *X-View*.

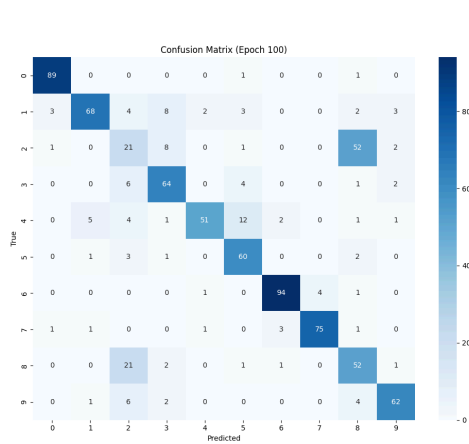
Tabela 9 – Mapeamento do conjunto de 10 ações.

ID	Ação
0	Pegar (ou "apanhar algo")
1	Arremessar
2	Bater palmas
3	Acenar com a mão
4	Chutar algo
5	Colocar a mão no bolso
6	Pular com um pé só
7	Saltar (pular para cima)
8	Esfregar as mãos
9	Cruzar as mãos à frente (gesto de "pare")

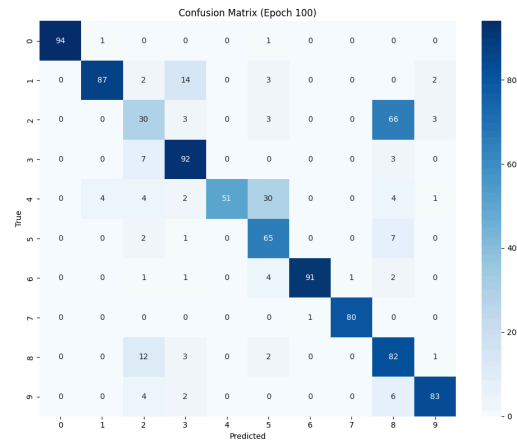
Ao analisar as matrizes de confusão, nota-se que, em todos os cenários, a maior taxa de erros ocorreu entre as ações de ID 2 e 8, que representam as ações de bater palmas e esfregar as mãos, respectivamente. Isso se dá por duas principais razões: a primeira é que as duas ações são semelhantes, o que pode acarretar uma baixa variabilidade entre as duas classes; a segunda está relacionada com a limitação no modelo de representação utilizada do esqueleto, que não traz informações minuciosas dos pés e das mãos.

Os resultados evidenciaram uma clara vantagem dos métodos que utilizam ST-GCN em contraste com os métodos baseados em 1D-CNN e informações de movimento e angulação dos membros, mostrando que o uso desse tipo de informação, limitado apenas aos membros como braços e pernas, pode não ser suficiente, principalmente para ações mais complexas que necessitam de informações detalhadas de pés e mãos para serem reconhecidas.

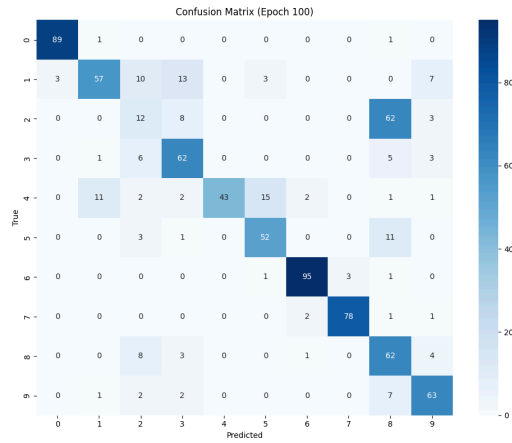
Figura 29 – Matrizes de confusão geradas pelos modelos ST-GCN utilizando poses 2D.



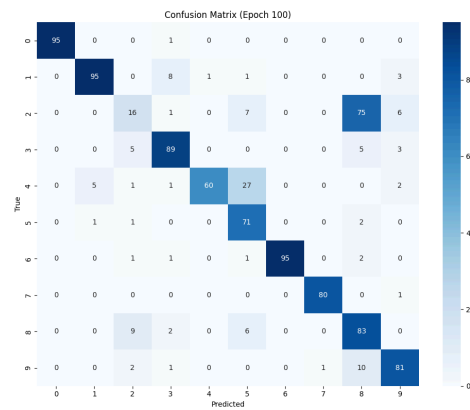
(a) ST-GCN com projeções RGB em 2D, protocolo X-Sub



(b) ST-GCN com projeções RGB em 2D, protocolo X-View



(c) ST-GCN com projeções Depth em 2D, protocolo X-Sub



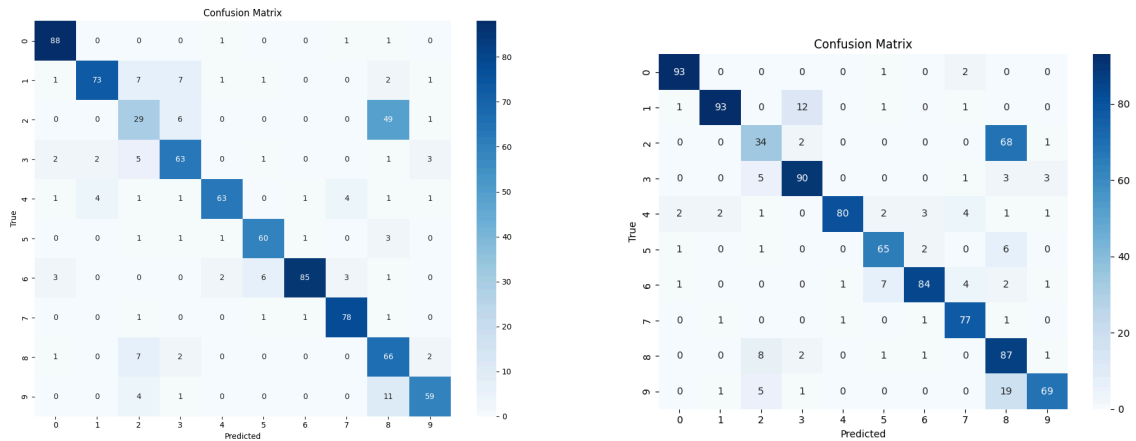
(d) ST-GCN com projeções Depth em 2D, protocolo X-View

Fonte: Elaborada pelo Autor.

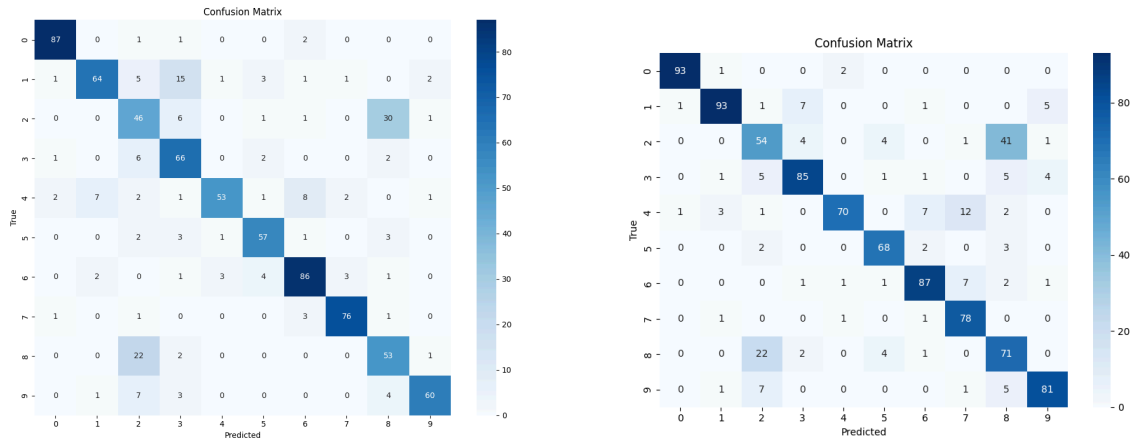
Tendo em vista o melhor desempenho obtido no conjunto de 10 ações pelo modelo ST-GCN, este foi selecionado para comparação com outros métodos do estado da arte encontrados na literatura, como os propostos por Luvizon et al. (2019b) e Ye et al. (2024). Para viabilizar uma comparação justa, utilizou-se o subconjunto de 40 ações adotado por trabalhos anteriores.

A variante aprimorada do ST-GCN, a ST-GCNMB, não foi incluída nesse experimento devido ao seu elevado custo computacional e à demanda significativamente maior por memória durante o treinamento e a inferência. Essas características tornam sua aplicação inviável em conjuntos de dados maiores, como o subconjunto de 40 ações, sem a utilização de recursos computacionais substancialmente mais robustos. Por esse motivo, optou-se por limitar sua avaliação ao cenário com 10 ações, o que permitiu explorar seu potencial com maior eficiência, sem comprometer a viabilidade experimental. Essa decisão visa garantir uma comparação justa

Figura 30 – Matrizes de confusão geradas pelos modelos ST-GCN utilizando poses 3D.



(a) ST-GCN com poses 3D obtidas pelo Kinect, protocolo X-Sub (b) ST-GCN com poses 3D obtidas pelo Kinect, protocolo X-View



(c) ST-GCN com poses 3D estimadas, protocolo X-Sub (d) ST-GCNBM com poses 3D estimadas, protocolo X-View

Fonte: Elaborada pelo Autor.

e reprodutível, mantendo a consistência com os recursos disponíveis para os demais modelos analisados.

A Tabela 10 apresenta as acurácias obtidas pelos métodos propostos neste trabalho, bem como por métodos correlatos.

Observa-se uma queda significativa no desempenho do modelo ST-GCN neste cenário, em comparação com as abordagens do estado da arte. Essa redução pode ser atribuída à dificuldade do modelo em distinguir ações que dependem fortemente de informações detalhadas sobre as articulações das mãos.

A abordagem proposta prioriza uma solução de baixo custo ao eliminar a necessidade do sensor *Kinect*, tornando-se, assim, dependente da saída de estimadores de pose 3D. Conse-

Tabela 10 – Acurácias obtidas por métodos estado-da-arte e pelo método ST-GCN proposto neste trabalho, no conjunto de 40 ações cotidianas do conjunto de dados NTU RGB+D.

Modelo	Tipo de Dado	X-Sub	X-View
Luvizon et al. (2019b)	3D Kinect	89.9%	-
Ye et al. (2024)	3D Kinect	89.5%	93.5%
<i>RGBFormer</i> (SHI et al., 2023)	3D Kinect	91.1%	-
<i>ST-GCN</i> (YAN et al., 2018)	3D Kinect	81.5%	88.3%
<i>2S-EGCN</i> (WANG et al., 2022a)	3D Kinect	89.1%	95.5%
<i>1s-Hybridnet</i> (YANG et al., 2023)	3D Kinect	88.3%	95.7%
<i>2s-Hybridnet</i> (YANG et al., 2023)	3D Kinect	90.4%	96.5%
<i>3s-Hybridnet</i> (YANG et al., 2023)	3D Kinect	91.4%	96.9%
ST-GCN	Projeção 2D RGB	59.56%	67.66%
ST-GCN	Projeção 2D Depth	60.47%	68.68%
ST-GCN	3D Kinect	66.15%	72.31%
ST-GCN	Poses 3D Estimadas	62.04%	66.68%

quentemente, o número de articulações disponíveis para o método é limitado pelo estimador adotado. Muitos estimadores de pose 3D em duas etapas contemporâneos, incluindo o modelo DUA, são avaliados ou treinados com base no conjunto de dados Human3.6M (IONESCU et al., 2013), que define um esqueleto com 16 articulações. Essa representação reduzida impõe limitações na descrição de regiões anatomicamente densas, como mãos e pés, o que pode prejudicar a acurácia da classificação em ações que envolvem movimentos sutis nessas áreas, além de introduzir ambiguidades na interpretação das ações no conjunto de dados.

Observa-se, entretanto, que mesmo neste conjunto de dados adverso para o modelo proposto, o seu desempenho foi superior quando se utilizaram as poses 3D estimadas comparando-se com as projeções 2D, mostrando que o uso de dados 3D favorece o reconhecimento mais preciso de ações humanas.

6 Conclusão

Este trabalho investigou a viabilidade de realizar o reconhecimento de ações humanas em 3D a partir de vídeos monoculares RGB, por meio da integração entre estimativas de pose 3D e modelos de reconhecimento baseados em estruturas esqueléticas. A proposta fundamenta-se no uso da técnica de estimação de pose 3D DUA, combinada com modelos de aprendizado profundo, como redes convolucionais unidimensionais (1D-CNN) e redes convolucionais em grafos espaço-temporais (ST-GCN), de forma a permitir a execução da tarefa sem a dependência de sensores de profundidade, como o *Kinect*.

Os experimentos realizados com o conjunto de dados NTU RGB+D demonstraram que a utilização de poses 3D estimadas, mesmo com limitações impostas por modelos simplificados de esqueleto, contribui positivamente para a acurácia do reconhecimento de ações, especialmente em comparação com abordagens baseadas exclusivamente em projeções 2D. O desempenho superior da configuração DUA-GCN frente às demais alternativas avaliadas, no subconjunto de 10 ações, reforça a relevância da dimensão adicional introduzida pela representação tridimensional, ainda que derivada de imagens monoculares.

A análise comparativa com métodos do estado da arte indicou que, embora a abordagem proposta apresente desempenho competitivo em cenários reduzidos, sua eficácia diminui ao lidar com conjuntos de ações mais amplos, como no caso das 40 classes do NTU RGB+D. Essa redução de desempenho pode ser atribuída à expressividade limitada da topologia esquelética utilizada, composta por 16 articulações, o que afeta negativamente a modelagem de movimentos sutis, especialmente aqueles envolvendo extremidades como mãos e pés. Essa limitação estrutural sugere que avanços em estimadores de pose que forneçam representações esqueléticas mais densas podem beneficiar diretamente o reconhecimento de ações em contextos complexos.

Outro aspecto relevante observado é a relação entre a qualidade da informação de entrada e a capacidade de generalização dos modelos. Enquanto o ST-GCN mostrou-se mais robusto na captação de padrões temporais e espaciais em dados 3D, a abordagem baseada em 1D-CNN apresentou sensibilidade maior à forma de entrada, destacando a importância de métodos de pré-processamento e representação para a eficácia do reconhecimento. Ainda que o método proposto não tenha superado os melhores resultados da literatura no cenário de 40 ações, os resultados obtidos validam o *pipeline* integrado como uma alternativa funcional e de menor custo computacional e estrutural.

A inclusão de uma variante do ST-GCN baseada em movimento, que incorpora informações geométricas relativas, como ângulos representados por quatérnios e distâncias euclidianas entre articulações, demonstrou potencial para enriquecer a representação espacial e melhorar a capacidade discriminativa do modelo. Essa abordagem permitiu capturar relações mais precisas

entre segmentos corporais ao longo do tempo, resultando no melhor desempenho entre os modelos avaliados no subconjunto de 10 ações. No entanto, sua aplicação em cenários mais complexos, como o conjunto com 40 ações, foi limitada por dois fatores principais: o elevado custo computacional, que demanda consideravelmente mais memória e processamento do que o modelo padrão, e a ausência de articulações mais detalhadas, como as das mãos e pés, que são fundamentais para a correta distinção de diversas ações presentes nesse conjunto. Embora essa última limitação também se aplique ao ST-GCN original, o impacto no ST-GCNMB é potencialmente mais crítico, uma vez que sua representação baseada em movimento depende diretamente da precisão dessas articulações para extrair relações geométricas significativas. Por esses motivos, optou-se por restringir a avaliação dessa variante ao cenário com 10 ações, onde tais limitações são menos impactantes e os recursos computacionais necessários permanecem viáveis.

Dessa forma, conclui-se que a integração entre estimação de pose 3D monocular e modelos de reconhecimento de ações baseados em grafos constitui uma alternativa viável para cenários em que sensores dedicados não estão disponíveis. No entanto, a qualidade das poses estimadas permanece como um fator determinante no desempenho final. Trabalhos futuros podem considerar a utilização de estimadores com maior resolução esquelética, bem como a incorporação de mecanismos de atenção temporal e espacial para lidar com a ambiguidade presente em ações similares. Além disso, a investigação de estratégias de fusão entre múltiplas fontes de entrada, como fluxo óptico ou *embeddings* visuais, pode contribuir para a superação das limitações observadas neste estudo.

Referências

- AHAD, M. A. R. Introduction. In: _____. *Computer Vision and Action Recognition: A Guide for Image Processing and Computer Vision Community for Action Understanding*. Paris: Atlantis Press, 2011. p. 1–8. ISBN 978-94-91216-20-6. Available on: <https://doi.org/10.2991/978-94-91216-20-6_1>.
- ANGELINI, F.; FU, Z.; LONG, Y.; SHAO, L.; NAQVI, S. M. 2d pose-based real-time human action recognition with occlusion-handling. *IEEE Transactions on Multimedia*, v. 22, n. 6, p. 1433–1446, 2020.
- ARNAB, A.; DEGHANI, M.; HEIGOLD, G.; SUN, C.; LUČIĆ, M.; SCHMID, C. *ViViT: A Video Vision Transformer*. 2021.
- BERTASIUS, G.; WANG, H.; TORRESANI, L. *Is Space-Time Attention All You Need for Video Understanding?* 2021.
- BLANK, M.; GORELICK, L.; SHECHTMAN, E.; IRANI, M.; BASRI, R. Actions as space-time shapes. In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. [S.l.: s.n.], 2005. v. 2, p. 1395–1402 Vol. 2.
- BUTTERWORTH, S. On the Theory of Filter Amplifiers. *Experimental Wireless & the Wireless Engineer*, v. 7, p. 536–541, out. 1930.
- CAMARENA, F.; GONZALEZ-MENDOZA, M.; CHANG, L.; CUEVAS-ASCENCIO, R. An overview of the vision-based human action recognition field. *Mathematical and Computational Applications*, v. 28, n. 2, 2023. ISSN 2297-8747. Available on: <<https://www.mdpi.com/2297-8747/28/2/61>>.
- CAO, Z.; HIDALGO, G.; SIMON, T.; WEI, S.-E.; SHEIKH, Y. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.
- CHEN, C.; JAFARI, R.; KEHTARNAVAZ, N. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: *2015 IEEE International Conference on Image Processing (ICIP)*. [S.l.: s.n.], 2015. p. 168–172.
- CHEN, Y.; TIAN, Y.; HE, M. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, v. 192, p. 102897, mar. 2020. ISSN 10773142.
- CIPTADI, A.; GOODWIN, M. S.; REHG, J. M. Movement pattern histogram for action recognition and retrieval. In: FLEET, D.; PAJDLA, T.; SCHIELE, B.; TUYTELAARS, T. (Ed.). *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014. p. 695–710. ISBN 978-3-319-10605-2.
- DANG, Q.; YIN, J.; WANG, B.; ZHENG, W. Deep learning based 2D human pose estimation: A survey. *Tsinghua Science and Technology*, v. 24, n. 6, p. 663–676, 2019. ISSN 18787606.
- DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S. et al. An

image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

DU, Y.; WANG, W.; WANG, L. Hierarchical recurrent neural network for skeleton based action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2015.

DUAN, H.; ZHAO, Y.; CHEN, K.; LIN, D.; DAI, B. Revisiting skeleton-based action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. p. 2969–2978.

GANIN, Y.; USTINOVA, E.; AJAKAN, H.; GERMAIN, P.; LAROCHELLE, H.; LAVIOLETTE, F.; MARCHAND, M.; LEMPITSKY, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, JMLR.org, v. 17, n. 1, p. 2096–2030, jan. 2016. ISSN 1532-4435.

GKALELIS, N.; KIM, H.; HILTON, A.; NIKOLAIDIS, N.; PITAS, I. The i3dpost multi-view and 3d human action/interaction database. In: *2009 Conference for Visual Media Production*. [S.l.: s.n.], 2009. p. 159–168.

HAN, S.; ACHAR, M.; LEE, S.; PEÑA-MORA, F. Empirical assessment of a rgb-d sensor on motion capture and action recognition for construction worker monitoring. *Visualization in Engineering*, v. 1, n. 6, 2013. Available on: <<http://www.viejournal.com/content/1/1/6>>.

HARTLEY, R.; ZISSERMAN, A. *Multiple View Geometry in Computer Vision*. 2. ed. [S.l.]: Cambridge University Press, 2004.

IONESCU, C.; PAPAVA, D.; OLARU, V.; SMINCHISESCU, C. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 36, n. 7, p. 1325–1339, 2013.

JAIN, A.; TOMPSON, J.; LECUN, Y.; BREGLER, C. Modep: A deep learning framework using motion features for human pose estimation. In: CREMERS, D.; REID, I.; SAITO, H.; YANG, M.-H. (Ed.). *Computer Vision – ACCV 2014*. Cham: Springer International Publishing, 2015. p. 302–315. ISBN 978-3-319-16808-1.

JANGUA, D.; MARANA, A. A new method for gait recognition using 2d poses. In: SBC. *Anais do XVI Workshop de Visão Computacional*. [S.l.], 2020. p. 69–74.

JANGUA, D. R. S.; MARANA, A. N. Gait recognition using 2d poses. In: *Edição Especial REIC: Artigos do 40º Concurso de Trabalhos de Iniciação Científica do CSBC (CTIC/CSBC)*. [S.l.: s.n.], 2021. v. 19, n. 2.

KAID, A. E.; BAÏNA, K. A systematic review of recent deep learning approaches for 3d human pose estimation. *Journal of Imaging*, v. 9, n. 12, p. 275, dez. 2023.

KANAZAWA, A.; BLACK, M. J.; JACOBS, D. W.; MALIK, J. End-to-end recovery of human shape and pose. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 7122–7131.

KARIM, F.; MAJUMDAR, S.; DARABI, H.; HARFORD, S. Multivariate lstm-fcns for time series classification. *Neural Networks*, v. 116, p. 237–245, 2019. ISSN 0893-6080. Available on: <<https://www.sciencedirect.com/science/article/pii/S0893608019301200>>.

KE, Q.; BENNAMOUN, M.; AN, S.; SOHEL, F.; BOUSSAID, F. A new representation of skeleton sequences for 3d action recognition. *IEEE*, jul 2017. Available on: <<https://doi.org/10.1109/%2Fcvpr.2017.486>>.

KENDALL, A.; GAL, Y.; CIPOLLA, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 7482–7491.

KIM, T. S.; REITER, A. *Interpretable 3D Human Action Analysis with Temporal Convolutional Networks*. 2017.

KOCABAS, M.; ATHANASIOU, N.; BLACK, M. J. Vibe: Video inference for human body pose and shape estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 5253–5263, 2020.

KONG, Y.; FU, Y. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, Springer, v. 130, n. 5, p. 1366–1401, 2022.

KREISS, S.; BERTONI, L.; ALAHI, A. Pifpaf: Composite fields for human pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 11977–11986.

LI, M.; CHEN, S.; CHEN, X.; ZHANG, Y.; WANG, Y.; TIAN, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. p. 3590–3598.

LI, S.; CHAN, A. B. 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. In: CREMERS, D.; REID, I.; SAITO, H.; YANG, M.-H. (Ed.). *Computer Vision – ACCV 2014*. Cham: Springer International Publishing, 2015. v. 9004, p. 332–347. ISBN 978-3-319-16808-1. Series Title: Lecture Notes in Computer Science.

LIN, T.; WANG, Y.; LIU, X.; QIU, X. *A Survey of Transformers*. arXiv, 2021. Available on: <<https://arxiv.org/abs/2106.04554>>.

LIU, J.; SHAHROUDY, A.; XU, D.; WANG, G. *Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition*. 2016.

LIU, W.; BAO, Q.; SUN, Y.; MEI, T. Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 55, n. 4, nov 2022. ISSN 0360-0300. Available on: <<https://doi.org/10.1145/3524497>>.

LOPER, M. M.; MAHMOOD, N.; ROMERO, J.; PONS-MOLL, G.; BLACK, M. J. Smpl: A skinned multi-person linear model. In: ACM NEW YORK, NY, USA. *ACM transactions on graphics (TOG)*. [S.l.], 2015. v. 34, n. 6, p. 1–16.

LUVIZON, D. C.; TABIA, H.; PICARD, D. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, v. 85, p. 15–22, dez. 2019. ISSN 00978493.

LUVIZON, D. C.; TABIA, H.; PICARD, D. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *CoRR*, abs/1912.08077, 2019. Available on: <<http://arxiv.org/abs/1912.08077>>.

MANESCO, J. R. R.; BERRETTI, S.; MARANA, A. N. Dua: a domain-unified approach for cross-dataset 3d human pose estimation. *Sensors*, MDPI, v. 23, n. 17, p. 7312, 2023.

MARTINEZ, J.; HOSSAIN, R.; ROMERO, J.; LITTLE, J. J. A simple yet effective baseline for 3d human pose estimation. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2017. p. 2659–2668.

MAZZIA, V.; ANGARANO, S.; SALVETTI, F.; ANGELINI, F.; CHIABERGE, M. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, v. 124, p. 108487, 2022. ISSN 0031-3203. Available on: <<https://www.sciencedirect.com/science/article/pii/S0031320321006634>>.

MORENO-NOGUER, F. 3d human pose estimation from a single image via distance matrix regression. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.]: IEEE, 2017. p. 2823–2832.

NIBALI, A.; HE, Z.; MORGAN, S.; PRENDERGAST, L. Numerical Coordinate Regression with Convolutional Neural Networks. *arXiv:1801.07372 [cs]*, maio 2018. ArXiv: 1801.07372.

PAPADOPOULOS, K.; GHORBEL, E.; AOUADA, D.; OTTERSTEN, B. *Vertex Feature Encoding and Hierarchical Temporal Modeling in a Spatial-Temporal Graph Convolutional Network for Action Recognition*. 2019.

PAPANDREOU, G.; ZHU, T.; CHEN, L.-C.; GIDARIS, S.; TOMPSON, J.; MURPHY, K. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: FERRARI, V.; HEBERT, M.; SMINCHISESCU, C.; WEISS, Y. (Ed.). *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018. p. 282–299. ISBN 978-3-030-01264-9.

PATRIKAR, D. R.; PARATE, M. R. Anomaly detection using edge computing in video surveillance system: review. *International Journal of Multimedia Information Retrieval*, Springer, v. 11, n. 1, p. 85–110, 2022.

PENG, W.; SHI, J.; XIA, Z.; ZHAO, G. *Mix Dimension in Poincaré Geometry for 3D Skeleton-based Action Recognition*. 2020.

PLIZZARI, C.; CANNICI, M.; MATTEUCCI, M. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, Elsevier BV, v. 208-209, p. 103219, jul 2021. Available on: <<https://doi.org/10.1016%2Fj.cviu.2021.103219>>.

RAPCZYŃSKI, M.; WERNER, P.; HANDRICH, S.; AL-HAMADI, A. A baseline for cross-database 3d human pose estimation. *Sensors*, MDPI, v. 21, n. 11, p. 3769, 2021.

RHODIN, H.; SPÖRRI, J.; KATIRCIOGLU, I.; CONSTANTIN, V.; SALZMANN, M.; FUA, P. Learning monocular 3d human pose estimation from multi-view images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 8437–8446.

RIZVI, S. M. H. Time series deep learning for robust steady-state load parameter estimation using 1d-cnn. *Arabian Journal for Science and Engineering*, Springer, v. 47, n. 3, p. 2731–2744, 2022.

SÁRÁNDI, I.; HERMANS, A.; LEIBE, B. Learning 3d human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. [S.l.: s.n.], 2023. p. 2956–2966.

SCHULDT, C.; LAPTEV, I.; CAPUTO, B. Recognizing human actions: a local svm approach. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*. [S.l.: s.n.], 2004. v. 3, p. 32–36 Vol.3.

SHAHROUDY, A.; LIU, J.; NG, T.-T.; WANG, G. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 1010–1019.

SHAIKH, M. B.; CHAI, D.; ISLAM, S. M. S.; AKHTAR, N. From cnns to transformers in multimodal human action recognition: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications*, Association for Computing Machinery (ACM), v. 20, n. 8, p. 1–24, jul. 2024. ISSN 1551-6865. Available on: <<http://dx.doi.org/10.1145/3664815>>.

SHI, J.; ZHANG, Y.; WANG, W.; XING, B.; HU, D.; CHEN, L. A novel two-stream transformer-based framework for multi-modality human action recognition. *Applied Sciences*, v. 13, n. 4, 2023. ISSN 2076-3417. Available on: <<https://www.mdpi.com/2076-3417/13/4/2058>>.

SHI, L.; ZHANG, Y.; CHENG, J.; LU, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. p. 12018–12027.

SONG, Y.; ZHANG, Z.; SHAN, C.; WANG, L. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 45, p. 1474–1488, 2022.

STAMOU, G.; KRINIDIS, M.; LOUTAS, E.; NIKOLAIDIS, N.; PITAS, I. 2D and 3D Motion Tracking in Digital Video. In: *Handbook of Image and Video Processing*. [S.l.]: Elsevier, 2005. p. 491–XVIII. ISBN 978-0-12-119792-6.

SUN, Z.; KE, Q.; RAHMANI, H.; BENNAMOUN, M.; WANG, G.; LIU, J. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 45, n. 3, p. 3200–3225, 2023.

TANG, W.; LONG, G.; LIU, L.; ZHOU, T.; JIANG, J.; BLUMENSTEIN, M. Rethinking 1d-cnn for time series classification: A stronger baseline. *arXiv preprint arXiv:2002.10061*, Feb, p. 1–7, 2020.

ULHAQ, A.; AKHTAR, N.; POGREBNA, G.; MIAN, A. *Vision Transformers for Action Recognition: A Survey*. 2022.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. *Attention Is All You Need*. arXiv, 2017. Available on: <<https://arxiv.org/abs/1706.03762>>.

VEERIAH, V.; ZHUANG, N.; QI, G.-J. *Differential Recurrent Neural Networks for Action Recognition*. 2015.

WANG, Q.; ZHANG, K.; ASGHAR, M. A. Skeleton-based st-gcn for human action recognition with extended skeleton graph and partitioning strategy. *IEEE Access*, v. 10, p. 41403–41410, 2022.

WANG, X.; DAI, Y.; GAO, L.; SONG, J. Skeleton-based action recognition via adaptive cross-form learning. In: *Proceedings of the 30th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2022. (MM '22), p. 1670–1678. ISBN 9781450392037. Available on: <<https://doi.org/10.1145/3503161.3547811>>.

WANG, Z.; NIE, X.; QU, X.; CHEN, Y.; LIU, S. Distribution-aware single-stage models for multi-person 3d pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. p. 13096–13105.

WEINLAND, D.; RONFARD, R.; BOYER, E. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, v. 104, n. 2, p. 249–257, 2006. ISSN 1077-3142. Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour. Available on: <<https://www.sciencedirect.com/science/article/pii/S1077314206001081>>.

XIA, L.; CHEN, C.-C.; AGGARWAL, J. K. View invariant human action recognition using histograms of 3d joints. In: *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. [S.l.: s.n.], 2012. p. 20–27.

YAN, S.; XIONG, Y.; LIN, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. [S.l.]: AAAI Press, 2018. (AAAI'18/IAAI'18/EAAI'18). ISBN 978-1-57735-800-8.

YANG, H.; GU, Y.; ZHU, J.; HU, K.; ZHANG, X. Pgcn-tca: Pseudo graph convolutional network with temporal and channel-wise attention for skeleton-based action recognition. *IEEE Access*, IEEE, v. 8, p. 10040–10047, 2020.

YANG, W.; ZHANG, J.; CAI, J. et al. Hybridnet: Integrating gcn and cnn for skeleton-based action recognition. *Applied Intelligence*, Springer, v. 53, p. 574–585, 2023. Available on: <<https://doi.org/10.1007/s10489-022-03436-0>>.

YANG, Y.; RAMANAN, D. Articulated Human Detection with Flexible Mixtures of Parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 35, n. 12, p. 2878–2890, dez. 2013. ISSN 0162-8828, 2160-9292.

YE, M.; LI, W.; PENG, Y.; ZHU, J.; QIAN, C. Dual-stream transformer-gcn model with contextualized representations learning for monocular 3d human pose estimation. *arXiv preprint arXiv:2504.01764*, 2024. Available on: <<https://arxiv.org/abs/2504.01764>>.

ZHANG, J.; TU, Z.; YANG, J.; CHEN, Y.; YUAN, J. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. p. 13232–13242.

ZHENG, C.; ZHANG, S.; YU, G.; YUAN, J.; CHENG, K. Motionbert: A unified perspective on 2d and 3d human motion prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. p. 13198–13207.

ZHENG, P.; LI, S.; XIA, L.; WANG, L.; NASSEHI, A. A visual reasoning-based approach for mutual-cognitive human-robot collaboration. *CIRP Annals*, v. 71, n. 1, p. 377–380, 2022. ISSN 0007-8506. Available on: <<https://www.sciencedirect.com/science/article/pii/S000785062200066X>>.

ZUFFI, S.; FREIFELD, O.; BLACK, M. J. From Pictorial Structures to deformable structures. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI: IEEE, 2012. p. 3546–3553.