CENTRO UNIVERSITÁRIO FEI

ERIC DE BAÉRE GRASSL

ANÁLISE DE MÉTODOS DE RECONHECIMENTO DE COMPORTAMENTO HUMANO EM AMBIENTES DOMÉSTICOS UTILIZANDO SENSORES RGB+D

São Bernardo do Campo

2021

ERIC DE BAÉRE GRASSL

ANÁLISE DE MÉTODOS DE RECONHECIMENTO DE COMPORTAMENTO HUMANO EM AMBIENTES DOMÉSTICOS UTILIZANDO SENSORES RGB+D

Dissertação de Mestrado, apresentada ao Centro Universitário da FEI para obtenção do título de Mestre em Engenharia Elétrica. Orientado pelo Prof. Dr. Plinio Thomaz Aquino.

São Bernardo do Campo

2021

de Baére Grassl, Eric.
Análise de métodos de Reconhecimento de Comportamento Humano em Ambientes Domésticos utilizando sensores RGB+D / Eric de Baére Grassl. São Bernardo do Campo, 2021.
90 f. : il.

Dissertação - Centro Universitário FEI.
Orientador: Prof. Dr. Plinio Thomaz Aquino Junior.

1. Reconhecimento de ações humanas. 2. Kinect. 3. Graph Convolutional Networks. 4. Shift-GCN. 5. MS-G3D. I. Thomaz Aquino Junior, Plinio, orient. II. Título.

Elaborada pelo sistema de geração automática de ficha catalográfica da FEI com os dados fornecidos pelo(a) autor(a).

centro universitário FEI

APRESENTAÇÃO DE DISSERTAÇÃO
ATA DA BANCA EXAMINADORA

Programa de Pós-Graduação Stricto Sensu em Engenharia Elétrica

Mestrado

PGE-10

Aluno: Eric de Baére Grassl **Matrícula:** 118306-0

Título do Trabalho: Análise de métodos de reconhecimento de comportamento humano em ambientes domésticos utilizando sensores RGB+D.

Área de Concentração: Inteligência Artificial Aplicada à Automação e Robótica

Orientador: Prof. Dr. Plinio Thomaz Aquino Junior

Data da realização da defesa: 22/02/2021

ORIGINAL ASSINADA

Avaliação da Banca Examinadora:

A banca foi realizada no dia 22 de Fevereiro às 14:00 horas, e se iniciou com a apresentação do aluno que demonstrou fluência no tema do trabalho. Na sequência a arguição ocorreu com indicações de melhorias no texto, e questionamentos técnicos sobre a execução do trabalho, onde o aluno respondeu a todas as questões de forma adequada demonstrando conhecimento pleno do tema. A aprovação foi por unanimidade.

São Bernardo do Campo, 22 / 02 / 2021.

MEMBROS DA BANCA EXAMINADORA	
Prof. Dr. Plinio Thomaz Aquino Junior	Ass.: _____
Paulo Roberto Ferreira Júnior	Ass.: _____
Prof.ª Dr.ª Esther Luna Colombini	Ass.: _____

A Banca Julgadora acima-assinada atribuiu ao aluno o seguinte resultado:

APROVADO ☒ REPROVADO ☐

VERSÃO FINAL DA DISSERTAÇÃO

APROVO A VERSÃO FINAL DA DISSERTAÇÃO EM QUE FORAM INCLUÍDAS AS RECOMENDAÇÕES DA BANCA EXAMINADORA

_____

Aprovação do Coordenador do Programa de Pós-graduação

_____

Prof. Dr. Carlos Eduardo Thomaz

Dedicado aos meus pais que se esforçaram tanto para me dar acesso à educação e a meu orientador que batalhou comigo por este projeto.

AGRADECIMENTOS

Aos professores e funcionários que compõem o Programa de Mestrado em Inteligência Artificial Aplicada à Automação e Robótica, em especial ao Prof. Dr. Plinio Thomaz Aquino Junior, e a todos os companheiros de pesquisa, companheiros de curso, membros da família e amigos que se envolveram e compartilharam parte desta trajetória da minha vida.

“Os que se encantam com a prática sem a ciência são como os timoneiros que entram no navio sem timão nem bússola, nunca tendo certeza do seu destino.”

Leonardo da Vinci

RESUMO

O trabalho proposto busca realizar uma revisão bibliográfica para levantar as bases de dados e reconhecedores de ações estado da arte na literatura recente, além de selecionar dois reconhecedores para compará-los. De acordo com o estudo realizado, os reconhecedores baseados em esqueletos 3D se mostraram promissores e os dois com maior desempenho foram selecionados: MS-G3D e Shift-GCN. Por outro lado, as bases de dados escolhidas foram a NTU RGB+D 120, com 11480 amostras e 120 classes, e a NTU RGB+D HOME, obtida por um subconjunto da NTU RGB+D 120 de 23784 amostras e 25 classes com o foco em ações realizadas em ambientes domésticos ou relacionadas à saúde. Assim, 12 configurações dos reconhecedores selecionados foram criadas e comparadas nas bases de dados propostas pelos métodos *cross-subject* e *cross-setup*, onde todas mostraram desempenhos estado da arte de até 85,80% para a base de dados NTU RGB+D 120 e até 84,32% para a NT RGB+D HOME. Além disso, foi possível observar que os reconhecedores baseados em esqueletos 3D são mais robustos em relação a ambientes e perspectivas diferentes e que as informações das arestas entre as juntas ou as movimentação entre as juntas dos esqueletos 3D em *frames* adjacentes resultaram em um maior desempenho dos reconhecedores. Por último, é possível observar que a fusão dos resultados de reconhecedores treinados com informações diferentes pode trazer um aumento considerável de seus desempenhos. Espera-se que o estudo realizado neste trabalho pode ser utilizado para guiar outros pesquisadores para a seleção de bases de dados e reconhecedores de ações, além de auxiliar a equipe do programa RoboFEI@HOME a incrementar a plataforma robótica HERA com capacidades de reconhecimento de comportamento humano.

Palavras-chave: Reconhecimento de ações humanas. Kinect. *Graph Convolutional Networks*. MS-G3D. Shift-GCN. NTU RGB+D 120. NTU RGB+D HOME.

ABSTRACT

The purpose of this project is to review the bibliography in order to understand the characteristics of the state-of-the-art databases and recognition methods for human action and to evaluate two of the studied methods. According to the recent literature, the 3D skeleton based methods have shown promising results, thus the two methods with better performance were selected: MS-G3D and Shift-GCN. On the other hand, to train these methods, the NTU RGB+D 120 database, composed by 114480 samples from 120 classes, and the NTU RGB+D HOME, composed by 23784 samples and 25 classes from a subset of the NTU RGB+D 120 database, where selected. Twelve configurations of the MS-G3D and Shift-GCN where created and evaluated in the proposed databases by the cross-subject and cross-setup methods where state-of-the-art performance was achieved with results up to 85,80% and 84,32% for the NTU RGB+D 120 and NTU RGB+D HOME databases, respectively. Furthermore, the 3D skeleton bases methods achieved a better performance on the cross-setup evaluation where different environments and view points where tested. Also, the configurations trained with bone information and joint motion information presented higher performance. At last, the results indicate that fusion from models trained with different information such as joint, bone, joint motion and bone motion information can achieve a better recognition performance. It is expected that this project may be used by other researchers as a guide in the process of choosing database or methods for human action recognition, and also help the RoboFEI@HOME team to increment their robot platform named HERA with human action recognition capabilities.

Keywords: Human Action Recognition. Kinect. Graph Convolutional Networks. MS-G3D. Shift-GCN. NTU RGB+D 120. NTU RGB+D HOME.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de uma Rede Neural Recorrente. Na direita, as operações realizadas pelo modelo foram expandidas em um grafo computacional para uma melhor visualização do fluxo de processamento de acordo com o tempo t. . 18

Figura 2 – Exemplo de uma Rede Neural Recorrente que mapeia uma entrada de tamanho N em uma saída de tamanho N. 19

Figura 3 – Operação de convolução com um *Kernel* 2×2 em uma entrada de dimensões 3×4. 20

Figura 4 – Comparação entre duas camadas de uma RNC, na parte superior, e uma RNA *fully connected*, na parte inferior. 21

Figura 5 – Exemplo de uma camada convolucional de uma RNC. 21

Figura 6 – Curva de uma função ReLU. 22

Figura 7 – Exemplo da operação de *pooling* de máximo em uma camada de uma RNC. O *pooling* repassará uma ativação para a próxima camada caso ocorra pelo menos uma ativação nos detectores da camada atual. 23

Figura 8 – Exemplo da arquitetura de uma RNC para o processamento de imagens. . . 23

Figura 9 – *Two-Stream CNN* utilizada para o reconhecimento de ações. 24

Figura 10 – Modelo de GCN utilizado por (KIPF; WELLING, 2016) com C entradas e F saídas. 25

Figura 11 – Comparação entre uma convolução 2D (formato matricial) e uma convolução em grafo. 26

Figura 12 – Exemplo de GCN utilizada para a classificação de documentos de textos em (DEFFERRARD; BRESSON; VANDERGHEYNST, 2016). 26

Figura 13 – Exemplo da operação realizada por uma ST-GCN para classificar ações em sequências de esqueletos 3D. 27

Figura 14 – ST-GCN utilizada para a predição de tráfego em (YU; YIN; ZHU, 2018). . 28

Figura 15 – Representação de um esqueleto 3D: (a) é a representação padrão utilizada por (YAN; XIONG; LIN, 2018) e (b) é a representação proposta pelo MS-G3D em (LIU, Z. et al., 2020) 29

Figura 16 – Arquitetura do reconhecedor de ações MS-G3D. 30

Figura 17 – Comparação entre a convolução em grafo comum (esquerda) e convolução em grafo em *shift* (direita). 31

Figura 18 – Exemplo de uma convolução em grafos em *shift*. 32
Figura 19 – Exemplo da validação para o método *multifold* com $K = 4$, onde todas as interações entre os $K - 1$ conjuntos de treinamento (em branco) e o conjunto validação (em verde) são testadas. 33
Figura 20 – Exemplo de partição realizada pelo método *cross-subject*. 34
Figura 21 – Exemplo de partição realizada pelo método *cross-view*. 35
Figura 22 – Exemplo de partição realizada pelo método *cross-setup*. 36
Figura 23 – Processo de reconhecimento de ações em vídeos RGB+D. 37
Figura 24 – Modalidades de dados para o reconhecimento de ações. 37
Figura 25 – *Frames* de profundidade em dois componentes: (a) imagem da componente horizontal e (b) imagem da componente vertical. 38
Figura 26 – Amostra gerada a partir de uma sequência esqueletos 2D ou 3D. 39
Figura 27 – Exemplo de um reconhecedor com dois fluxos de processamento. 40
Figura 28 – Modalidades de dados e classificadores utilizados pelos reconhecedores de ações estudados. 58
Figura 29 – Métodos de validação utilizados pelos reconhecedores estudados. 59
Figura 30 – Metodologia para a obtenção das bases NTU RGB+D120 e NTU HOME. . 61
Figura 31 – Sensor Microsoft Kinect V2. 61
Figura 32 – Perspectivas de coleta para cada *setup* na base de dados NTU RGB+D 120. Existem 5 perspectivas para cada *setup*, sendo elas: frontal, lateral direita 45º, lateral direita, lateral esquerda 45º e lateral esquerda. O voluntário está voltado para a câmera da esquerda em (a), enquanto ele se encontra voltado para a câmera da direita em (b). 62
Figura 33 – Arquivos da base de dados NTU RGB+D 120. 63
Figura 34 – As 25 juntas que são utilizadas para formar os esqueletos 3D na base de dados NTU RGB+D 120. 64
Figura 35 – Modelos de reconhecedores propostos. 68
Figura 36 – Matriz de confusão obtida pelo MS-G3D na validação *cross-subject* na base NTU RGB+D 120. 73
Figura 37 – Matriz de confusão obtida pelo Shift-GCN na validação *cross-subject* na base NTU RGB+D 120. 74
Figura 38 – Matriz de confusão obtida pelo MS-G3D na validação *cross-setup* na base NTU RGB+D 120. 75

Figura 39 – Matriz de confusão obtida pelo Shift-GCN na validação *cross-setup* na base NTU RGB+D 120. 76
Figura 40 – Matriz de confusão obtida pelo MS-G3D na validação *cross-subject* na base NTU RGB+D HOME. 77
Figura 41 – Matriz de confusão obtida pelo Shift-GCN na validação *cross-subject* na base NTU RGB+D HOME. 78
Figura 42 – Matriz de confusão obtida pelo MS-G3D na validação *cross-setup* na base NTU RGB+D HOME. 79
Figura 43 – Matriz de confusão obtida pelo Shift-GCN na validação *cross-setup* na base NTU RGB+D HOME. 80

LISTA DE TABELAS

Tabela 1 – Comparação das Bases de Dados. As modalidades de dados são: RGB para imagens RGB, D para mapas de profundidade, 3DJ para juntas 3D e IR para infravermelho. 44

Tabela 2 – Descrição dos *setups* utilizados para a coleta dos vídeos da base NTU RGB+D 120. Vale notar que os *setups* 04 e 05 tem a mesma distância e altura da câmera, resultando nas mesmas perspectivas. 62

Tabela 4 – Comparação entre as bases de dados NTU RGB+D 120 e NTU RGB+D HOME. 64

Tabela 3 – Classes de ações selecionadas para a criação da base de dados NTU RGB+D HOME. 65

Tabela 5 – Resultados dos reconhecedores pelo método de validação *cross-subject* na base NTU RGB+D 120. 72

Tabela 6 – Resultados dos reconhecedores pelo método de validação *cross-setup* na base NTU RGB+D 120. 75

Tabela 7 – Resultados dos reconhecedores pelo método de validação *cross-subject* na base NTU RGB+D HOME. 77

Tabela 8 – Resultados dos reconhecedores pelo método de validação *cross-setup* na base NTU RGB+D HOME. 79

LISTA DE ABREVIATURAS E SIGLAS

FEI	Fundação Educacional Inaciana Pe. Sabóia de Medeiros
GCN	Graph Convolutional Network
HERA	HOME ENVIRONMENT ROBOT ASSISTANT
LSTM	Long Short Term Mermory
MLP	Multi-layer Perceptron
MS-G3D	Multi-scale Graph 3D Convolutional Network
NTU	Nanyang Technological University
RGB+D	Red Green Blue + Depth
RNA	Rede Neural Artificial
RNC	Rede Neural Convolucional
RNR	Rede Neural Recorrente
ST-GCN	Spatial-Temporal Graph Convolutional Network
SVM	Support Vector Machine

SUMÁRIO

1 INTRODUÇÃO . . . 15
2 CONCEITOS FUNDAMENTAIS . . . 17
2.1 REDES NEURAIS RECORRENTES . . . 17
2.2 REDES NEURAIS CONVOLUCIONAIS . . . 19
2.2.1 ***Two-Stream Convolutional Neural Network*** . . . 23
2.3 GRAPH CONVOLUTIONAL NETWORKS . . . 24
2.3.1 ***Spatial-Temportal Graph Convolutional Neural Network*** . . . 27
2.3.2 MS-G3D . . . 28
2.3.3 Shift-GCN . . . 30
2.4 VALIDAÇÃO CRUZADA . . . 32
2.4.1 Validação cross-subject . . . 34
2.4.2 Validação cross-view . . . 34
2.4.3 Validação cross-setup . . . 35
2.5 RECONHECIMENTO DE AÇÕES EM VÍDEOS RGB+D . . . 36
2.5.1 MODALIDADE DE DADOS . . . 36
2.5.2 PRÉ-PROCESSAMENTO DAS AMOSTRAS . . . 38
2.5.3 RECONHECIMENTO . . . 39
3 REVISÃO BIBLIOGRÁFICA . . . 41
3.1 BASES DE DADOS RGB+D . . . 41
3.2 RECONHECIMENTO DE AÇÕES . . . 45
4 METODOLOGIA . . . 60
4.1 ETAPA I - BASES DE DADOS . . . 60
4.1.1 Geração da base de dados NTU HOME (I.A) . . . 60
4.1.2 Transformação das sequências de esqueletos 3D em grafos (I.B) . . . 65
4.1.2.1 Juntas 3D (I.B.1) . . . 66
4.1.2.2 Arestas (I.B.2) . . . 66
4.1.2.3 Movimentação (I.B.3) . . . 66
4.2 ETAPA II - RECONHECEDORES . . . 67
4.2.1 MS-G3D (II.A) . . . 67
4.2.2 Shift-GCN (II.B) . . . 69

5 EXPERIMENTOS 70
5.1 NTU RGB+D 120 70
5.2 NTU RGB+D HOME 71
6 RESULTADOS 72
6.1 NTU RGB+D 120 72
6.2 NTU RGB+D HOME 76
6.3 DISCUSSÃO 80
7 CONCLUSÃO 83
REFERÊNCIAS 85

1 INTRODUÇÃO

O reconhecimento de comportamento humano vem sendo extensamente explorado pela literatura, uma vez que é utilizado em diversas áreas tais como monitoramento por vídeo, esportes e assistência doméstica (WANG; HUYNH; KONIUSZ, 2019). Em 2019, a funcionalidade de detecção de queda foi disponibilizada para os dispositivos Apple Watch Series 4 ou superior por meio de seus sensores embarcados, onde um contato de emergência é acionado caso uma queda brusca seja detectada. Vale notar que esse recurso é automaticamente liberado caso o usuário tenha mais de 54 anos de idade. Além disso, uma base de dados com a motivação de realizar monitoramento médico de pessoas por vídeos RGB é proposta em Papakostas et al. (2016), sendo seu foco no monitoramento de idosos por meio de um robô de assistência doméstica.

Apesar de suas diversas aplicações, ainda há desafios a serem explorados nesta área. Por exemplo, velocidades de execução de movimentos variáveis e ações compostas por movimentos semelhantes podem reduzir o desempenho dos métodos de reconhecimento de comportamento humano. O último caso pode ser observado para as ações de beber água e atender o telefone, onde ambas são compostas pela movimentação de um objeto ao rosto de um indivíduo. Por outro lado, fatores tais como a perspectiva da câmera e a oclusão parcial do corpo humano também podem diminuir o desempenho do reconhecimento (WANG; HUYNH; KONIUSZ, 2019).

Existem três níveis semânticos para o reconhecimento do comportamento humano, sendo eles: gestos, ações e atividades. Neste trabalho, o nível semântico explorado é o de ações, que são movimentos compostos por um conjunto de gestos, tais como beber água, andar, sentar, cair e falar no telefone (VAQUETTE et al., 2017). Este trabalho tem como objetivo:

a) Realizar uma revisão bibliográfica na literatura recente para levantar quais são as características das bases de dados e reconhecedores de ações;
b) Selecionar dois reconhecedores promissores na literatura e compará-los em duas bases de dados volumosas.

De acordo com a literatura estudada, foi possível observar que os reconhecedores que processam esqueletos 3D apresentaram um desempenho promissor em trabalhos recentes, assim como é destacado em Ziyu Liu et al. (2020). Portanto, dentre os trabalhos estudados, dois métodos baseado em esqueletos 3D foram selecionados para sua avaliação, sendo eles o MS-G3D e o Shift-GCN, onde diversas configurações deles foram geradas e comparadas. Além

disso, as bases escolhidas para o treinamento desses reconhecedores foram a NTU RGB+D 120, com 114480 amostras e 120 classes de ações, e um subconjunto da mesma denominado NTU RGB+D HOME com 23784 amostras e 25 classes com o foco em ações que ocorrem em ambientes domésticos ou relacionadas à saúde. As contribuíções do estudo realizado neste trabalho são:

a) Os dados e resultados levantados podem ser utilizados por outros pesquisadores no entendimento das bases de dados e reconhecedores estado da arte;
b) Outros pesquisados podem utilizar este trabalho como base para a comparação de desempenhos de reconhecedores baseados em esqueletos 3D;
c) Este trabalho mostra que é possível aplicar os reconhecedores avaliados em aplicações de reconhecimento de ações humanas, incluindo a equipe RoboFEI@HOME que pode utilizá-los para incrementar a plataforma robótica HERA (AQUINO JUNIOR et al., 2019).

Os conceitos e teorias necessários para compreender a metodologia deste trabalho se encontram no Capítulo 2 e uma revisão bibliográfica é realizada no Capítulo 3 com o objetivo de levantar as características das bases de dados e reconhecedores de ações da literatura recente. O Capítulo 4 detalha a metodologia deste trabalho, onde as bases NTU RGB+D 120 e NTU RGB+D HOME foram escolhidas, além das 12 configurações dos reconhecedores MS-G3D e Shift-GCN. O protocolo dos experimentos realizados em ambas as bases se encontram no Capítulo 5. A apresentação, análise e discussão dos resultados dos experimentos executados se encontram no Capítulo 6 e, por último, a conclusão do trabalho é realizada no Capítulo 7.

2 CONCEITOS FUNDAMENTAIS

Este Capítulo tem como objetivo descrever os conceitos utilizados para o desenvolvimento da metodologia e experimentos deste trabalho, visando permitir ao leitor o entendimento dos mesmos. Uma vez que o trabalho proposto utiliza principalmente Redes Neurais Artificiais (RNAs), as principais RNAs utilizadas pela literatura para o reconhcimento de ações serão descritas nas próximas Seções. Assim, a Seção 2.1 descreve as Redes Neurais Recorrentes, a Seção 2.2 detalha as Redes Neurais Convolucionais e os métodos que a utilizam, a Seção 2.3 descreve as *Graph Convolutional Networks* e métodos derivados, os principais métodos de validação utilizados na literatura são descritos na Seção 2.4 e, por último, o problema do reconhecimento de ações humanas é descrito na Seção 2.5.

2.1 REDES NEURAIS RECORRENTES

Da mesma forma que a Rede Neural Convolucional é uma RNA especializada no processamento de informações em formato matricial como imagens, a Rede Neural Recorrente (RNR), proposta em Rumelhart, Hinton e Williams (1986), é uma RNA especializada no processamento de dados sequências tais como sequências de palavras ou séries temporais (GOODFELLOW; BENGIO; COURVILLE, 2016).

Uma característica importante da RNR é sua capacidade de processamento de entradas de tamanho variável, que é resultado de seu compartilhamento de parâmetros de processamento. Por exemplo, dado um conjunto de dados sequenciais $x^{(1)}, ..., x^{(T)}$ onde os parâmetros são gerados individualmente para cada valor de t, sendo $1 \leq t \leq T$, a RNR não seria capaz de processar entradas de tamanho variável, uma vez que os parâmetros para os valores de $t > T$ não existem.

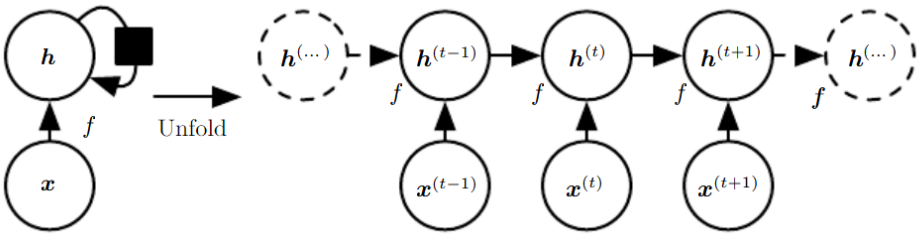
Assim, os neurônios artificiais de uma RNR são responsáveis por processar uma entrada, de acordo com o estado obtido pelo neurônio artificial anterior. A Figura 1 e a Equação 1 descrevem o processamento realizado por essa unidade, onde $h^{(t)}$ é o estado do neurônio no momento t e f é função de processamento do estado anterior $h^{(t-1)}$ e da entrada atual $x^{(t)}$, que pode ser expandida para a função $g^{(t)}$ (Equação 2). Uma vez que o processamento do sistema pode ser representado por uma função f compartilhada para todos os valores de t, a RNR torna-se independente da quantidade de dados de entrada $x^{(1)}$,...,$x^{(t)}$ e o modelo obtido é capaz de generalizar sequências de tamanhos que não foram utilizados em seu treinamento, além de

permitir o treinamento com uma quantidade menor de amostras em relação a um modelo sem o compartilhamento realizado pela função f.

$$h^{(t)} = f(h^{(t-1)},x^{(t)},\theta) \tag{1}$$

$$h^{(t)} = g^{(t)}(x^{(t)},x^{(t-1)},x^{(x-2)},...,x^{(2)},x^{(1)}) \tag{2}$$

Figura 1 – Exemplo de uma Rede Neural Recorrente. Na direita, as operações realizadas pelo modelo foram expandidas em um grafo computacional para uma melhor visualização do fluxo de processamento de acordo com o tempo t.



Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016).

A Figura 2 ilustra uma RNR que mapeia uma sequência de entrada de tamanho N em uma sequência de saída do mesmo tamanho. Esse modelo de RNA pode ser utilizado para a predição de palavras em uma frase, por exemplo. Assim, para cada valor de t, as Equações 3, 4, 5 e 6 são aplicadas para obter o estado atual $h^{(t)}$, a saída não-normalizada o e sequência de saída mapeada $\hat{y}$, sendo U, V e W as matrizes de pesos de conexões *input-to-hidden*, *hidden-to-output* e *hidden-to-hidden*, respectivamente, e, por último, b e c os vetores de *bias*.

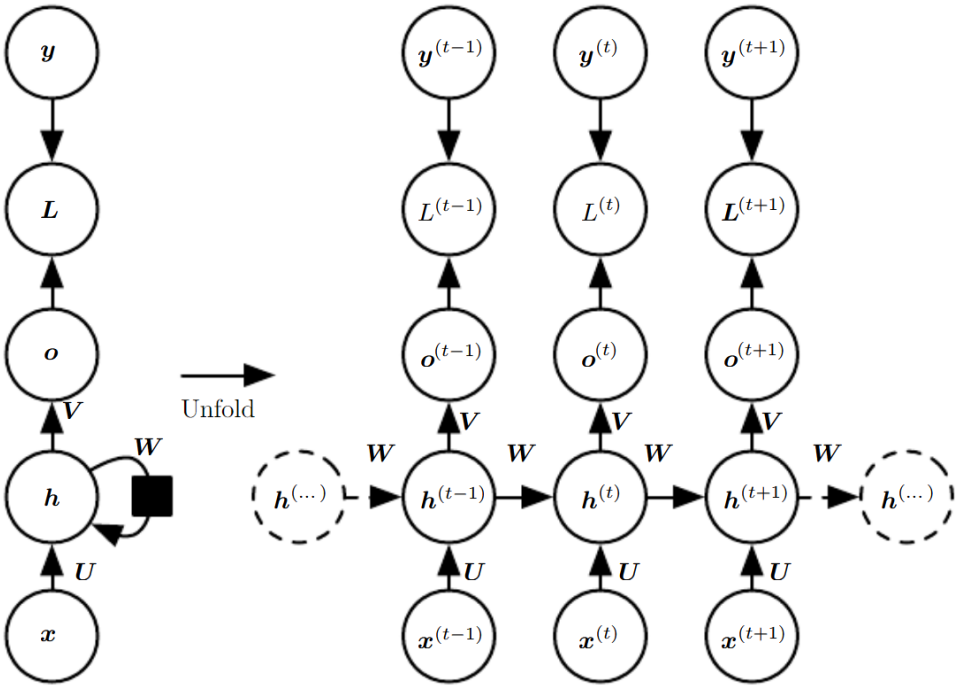
$$a^{(t)} = b + Wh^{(t-1)} + Ux^{(t)} \tag{3}$$

$$h^{(t)} = tanh(a^{(t)}) \tag{4}$$

$$o^{(t)} = c + Vh^{(t)} \tag{5}$$

$$\hat{y}^{(t)} = softmax(o^{(t)}) \tag{6}$$

Figura 2 – Exemplo de uma Rede Neural Recorrente que mapeia uma entrada de tamanho N em uma saída de tamanho N.



Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016).

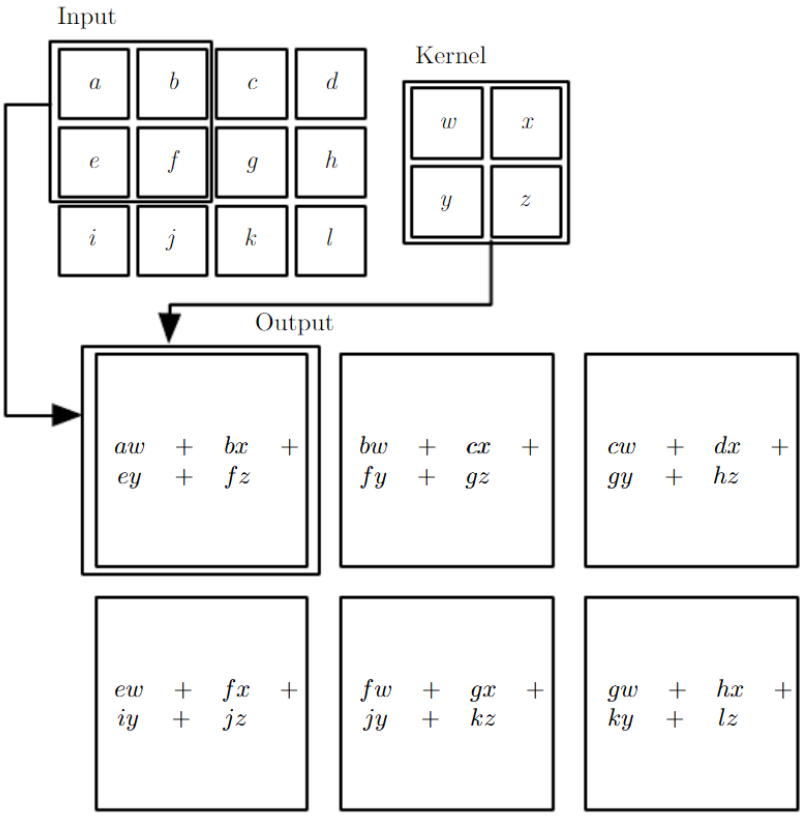
2.2 REDES NEURAIS CONVOLUCIONAIS

A Rede Neural Convolucional (RNC) é uma RNA proposta em Lecun et al. (1998) inspirada no córtex visual de animais, que visa reproduzir os padrões de conectividade de neurônios de animais por meio de operações de convolução. Além disso, seu aprendizado é supervisionado, sendo geralmente utilizado o algoritmo *Back Propagation*. Assim, a RNC pode ser definida como uma RNA onde uma ou mais camadas substituem cálculo de multiplicação matricial por convoluções especializadas no processamento de dados em formato matricial como imagens, por exemplo.

A operação de convolução é caracterizada pela sobreposição de um *Kernel* sobre todas as entradas da RNA e está descrita na Figura 3. Neste caso, a convolução é feita por um *Kernel* de tamanho 2×2 em uma entrada de tamanho 3×4. Observa-se que os valores do *Kernel* são os pesos que conectam Neurônios Artificiais de duas camadas e são compartilhados por todos os Neurônios Artificiais envolvidos na convolução. Assim, as conexões entre as camadas são esparsas resultando em uma maior capacidade de processamento de dados pela RNC. A Figura 4 compara as conexões entre camadas de uma RNC, no topo da imagem, com camadas de uma

RNA completamente conectada (*fully connected*), na parte inferior da imagem, onde há menos pesos no caso da RCN, resultando em uma conectividade esparsa. Por último, a RNC é capaz de generalizar dados automaticamente, sem pré-processamento da entrada, mesmo quando os dados forem transladados espacialmente.

Figura 3 – Operação de convolução com um *Kernel* 2×2 em uma entrada de dimensões 3×4.

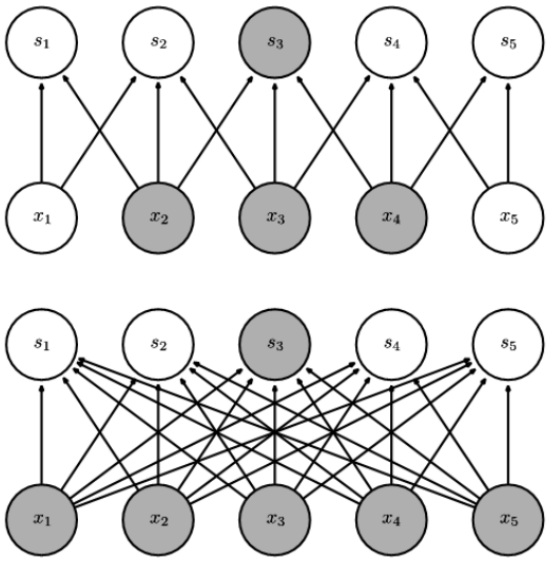


Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016).

Uma camada de uma RNC é geralmente composta por três procedimentos ilustrados pela Figura 5: convoluções, detecção e *pooling*. O processamento executado por uma camada se inicia as pela aplicação de transformações nos dados de entrada por meio de operações de convolução. Após a transformação dos dados, os mesmos são aplicados à um procedimento de detecção onde funções linearmente retificadas são aplicadas aos mesmos. Por ultimo, o procedimento de *pooling* é aplicado visando obter um resumo estatístico de vizinhanças dos valores obtidos (GOODFELLOW; BENGIO; COURVILLE, 2016).

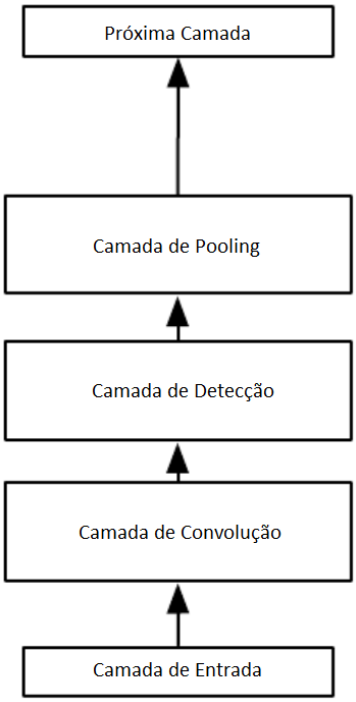
O procedimento de detecção responsável por obter características para a classificação dos dados de entrada e é geralmente obtido por meio da aplicação das entradas em funções *Rectified Linear Unit* (ReLU), que são funções de ativação linear que convergem de acordo

Figura 4 – Comparação entre duas camadas de uma RNC, na parte superior, e uma RNA *fully connected*, na parte inferior.



Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016).

Figura 5 – Exemplo de uma camada convolucional de uma RNC.

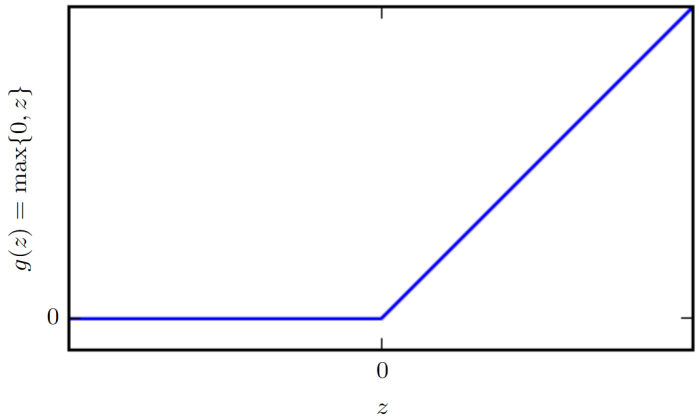


Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016).

com o gradiente descendente. A Equação 7 descreve a função e a Figura 6 descreve a curva da mesma.

$$f(x) = max(0,x) \tag{7}$$

Figura 6 – Curva de uma função ReLU.

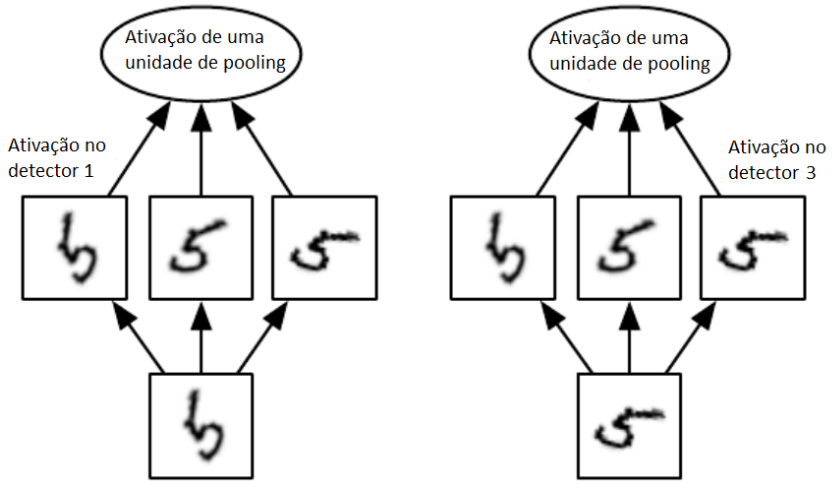


Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016).

O procedimento de *pooling* substitui a saída de um local de uma RNC por um resumo estatístico da mesma buscando reduzir a dimensionalidade de sua entrada e auxiliar a representação da mesma a ser invariante a translações e rotações. Ele pode ser realizado de forma local, através de um *kernel*, ou de forma global em sua entrada e ocorre por meio de operações matemáticas tais como média ou máxima. Essa operação é essencial para aumentar a precisão da RNC e auxiliar na classificação para entradas de diversos tamanhos. A Figura 7 ilustra um procedimento de *pooling* de máximo global para a detecção de dígitos feitos à mão, onde pode-se observar a invariância à translação e rotação, uma vez que a próxima camada será ativada caso qualquer um dos 3 detectores sejam ativados.

Para evitar o *overfitting* da RNC, o procedimento de *Dropout* pode ser utilizado. Esse procedimento evita com que os Neurônios Artificiais da RNC se especializem em reconhecer um determinado padrão com o objetivo de evitar *overfitting*. Esse procedimento consiste em desativar aleatoriamente uma porção dos Neurônios Artificiais da RNC, onde os neurônios artificiais ativos terão de substituir a função dos neurônios artificiais inativos, evitando a especialização dos mesmos. Por último, a Figura 8 mostra uma RNC composta pelas camadas e procedimentos anteriores, onde a classificação é obtida por uma última camada *fully connected* aplicada na saída da última camada convolucional.

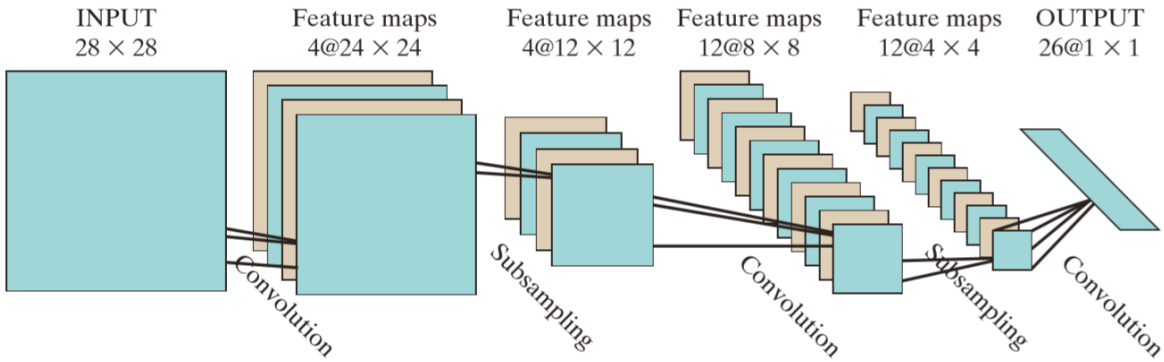
Figura 7 – Exemplo da operação de *pooling* de máximo em uma camada de uma RNC. O *pooling* repassará uma ativação para a próxima camada caso ocorra pelo menos uma ativação nos detectores da camada atual.



11

Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016).

Figura 8 – Exemplo da arquitetura de uma RNC para o processamento de imagens.



Fonte: (HAYKIN, 2009).

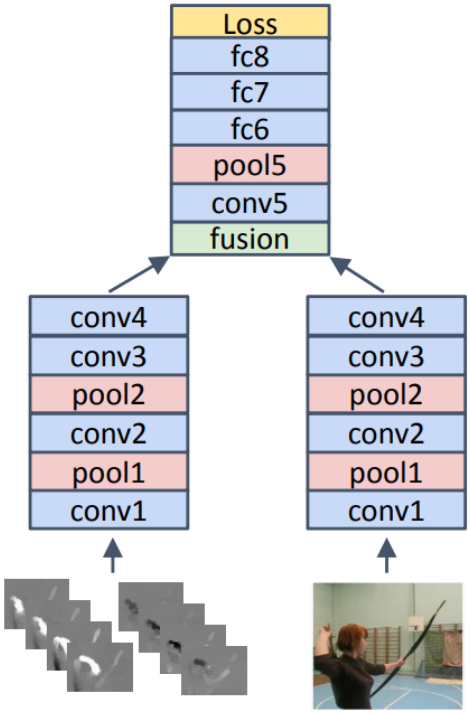
2.2.1 *Two-Stream Convolutional Neural Network*

O método *Two-Stream Convolutional Neural Network* (*Two-Stream* CNN), proposto em Simonyan e Zisserman (2014), é utilizado para realizar a análise estrutural e temporal de sua entrada, que é composta por um conjunto de *frames* RGB. Esse modelo é baseado na hipótese de que o córtex visual do cérebro humano é separado em dois fluxos: um de reconhecimento de objetos (Fluxo Ventral) e outro para o reconhecimento de movimentação (Fluxo Dorsal).

A Figura 9 ilustra um modelo de *Two-Stream CNN* utilizado para o reconhecimento de ações em uma sequência de vídeos RGB. O modelo é separado em duas RNCs diferentes, sendo uma para obter as dependências estruturais da entrada e outra para as dependências temporais.

No caso da RNC estrutural, os *frames* RGB são analisados de forma unitária e informações da cena são obtidas. Vale lembrar que essa RNC pode ser pré-treinada em outras bases de imagens tais como a ImageNet. Por outro lado, a RNC temporal analisa um conjunto de fluxos ópticos que representam a movimentação realizada pelos objetos e pela câmera nos *frames*. Neste caso, esta RNC necessita de bases de vídeos para seu treinamento, não podendo seguir a mesma abordagem da RNC estrutural. Por último, as saídas de ambas RNCs sofrem fusão e a classificação final é obtida.

Figura 9 – *Two-Stream CNN* utilizada para o reconhecimento de ações.



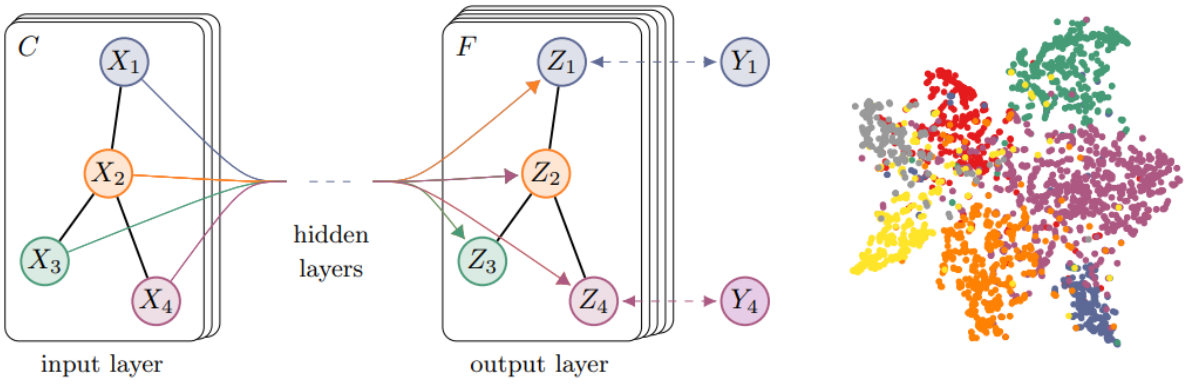
Fonte: (FEICHTENHOFER; PINZ; ZISSERMAN, 2016)

2.3 GRAPH CONVOLUTIONAL NETWORKS

RNAs que processam informações em estruturas de grafos foram inicialmente propostas por Gori, Monfardini e Scarselli (2005). Em seguida, trabalhos tais como Scarselli et al. (2009) e Yujia Li et al. (2016) utilizaram conceitos de RNRs (Seção 2.1) neste modelo. Por outro lado, o processamento de grafos por convoluções foi proposto em Bruna et al. (2013) e desenvolvido em trabalhos como Duvenaud et al. (2015) e Kipf e Welling (2016), onde uso em grafos volumosos foi permitido por meio da convolução com matrizes de pesos únicos por camada.

Este modelo (Figura 10), denominado *Graph Convolutional Network* (GCN), é utilizado neste trabalho para o processamento de esqueletos 3D formados por nós e vértices.

Figura 10 – Modelo de GCN utilizado por (KIPF; WELLING, 2016) com C entradas e F saídas.



(a) Graph Convolutional Network (b) Hidden layer activations

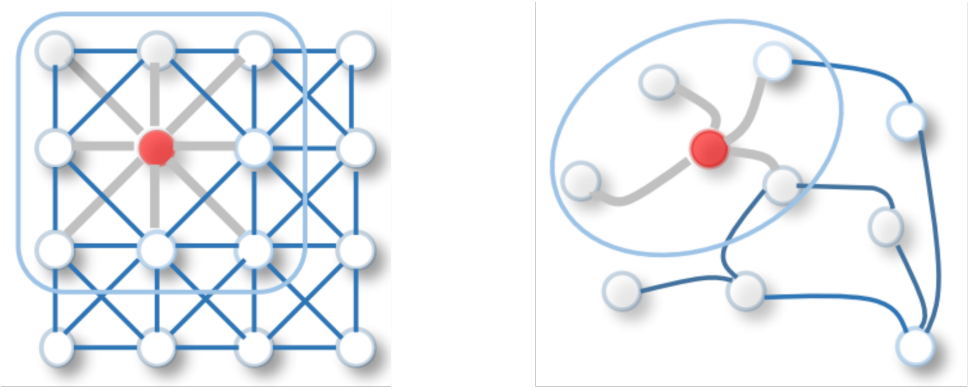
Fonte: (KIPF; WELLING, 2016).

Uma operação de convolução de uma camada da GCN busca processar informações estruturais de sua entrada. A Figura 11 ilustra a comparação entre uma convolução 2D em uma estrutura matricial e uma convolução em um grafo, onde é possível observar que a convolução 2D tem seu tamanho fixo, uma vez que ela é realizada sobre uma quantidade e ordem fixa de dados, enquanto a convolução em um grafo envolve nós que não estão em uma sequência ordenada e suas vizinhanças têm tamanhos variáveis de acordo com a quantidade de ligações existentes. Assim, para cada camada da GCN, a Equação 8 é aplicada para obter-se a saída da convolução, onde $\tilde{A}$ é a matriz de adjacência do grafo, $\tilde{A} = A + I$, $\tilde{D}$ é a matriz com o número de conexões recebidas por cada nó com $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$, $H^{(l)}$ a matriz de ativações da camada l, $W^{(l)}$ a matriz de pesos para a camada l e σ é a função de ativação. Vale notar que os pesos $W^{(l)}$ são inicializados aleatoriamente e aprendidos durante o processo de treinamento do modelo.

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{8}$$

Assim, uma CGN pode ser obtida por meio da utilização de duas ou mais camadas convolucionais, descritas no parágrafo anterior, onde um descritor para o grafo de entrada é gerado. Por exemplo, a Figura 12 descreve uma GCN proposta em Defferrard, Bresson e Vandergheynst (2016) com o objetivo de classificar documentos de texto. O modelo é composto por duas camadas convolucionais, uma camada de *pooling* e, por último, uma RNA MLP seguida por uma camada *softmax* para realizar a classificação da entrada.

Figura 11 – Comparação entre uma convolução 2D (formato matricial) e uma convolução em grafo.

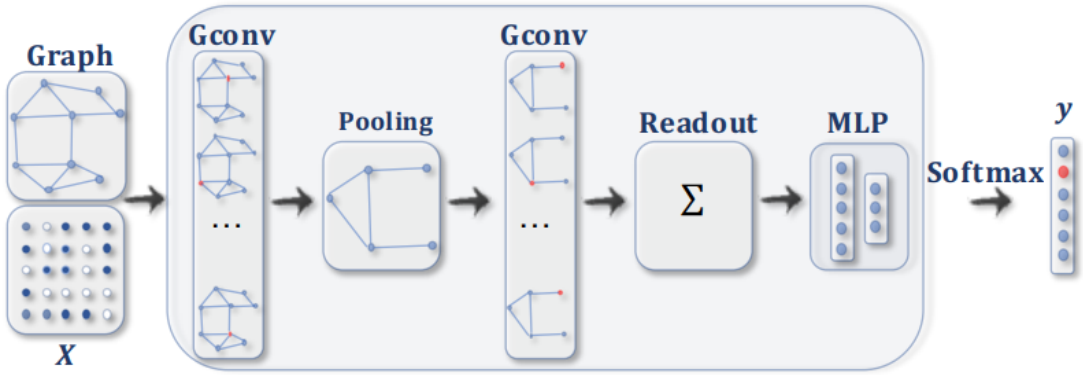


(a) Convolução 2D. Análogo a um grafo, cada pixel em uma imagem é considerado um nó onde suas vizinhança é determinada pelo tamanho do filtro. A convolução 2D utiliza a média ponderada dos valores do pixel vermelho e seus vizinhos. Os vizinhos de um nó são ordenados e têm um tamanho fixo.

(b) Convolução em Grafo. Para obter uma representação do nó vermelho, uma solução para a operação de convolução é obter a média ponderada do nó vermelho e de seus vizinhos. Diferente dos dados de uma imagem, os vizinhos de um nó não são ordenados e tem tamanho variável.

Fonte: (WU et al., 2019).

Figura 12 – Exemplo de GCN utilizada para a classificação de documentos de textos em (DEFFERRARD; BRESSON; VANDERGHEYNST, 2016).



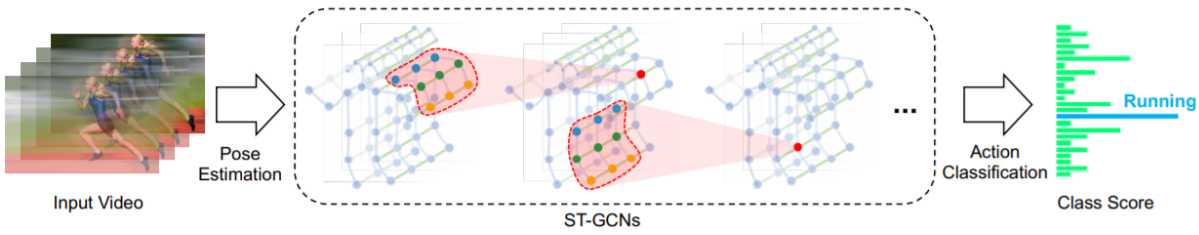
Fonte: (WU et al., 2019)

Por último, as possíveis aplicações de uma GCN envolvem problemas que podem ser convertidos para uma classificação de nós de um grafo tais como redes de citação de artigos (KIPF; WELLING, 2016), análise de documentos de texto (DEFFERRARD; BRESSON; VANDERGHEYNST, 2016) e análise de moléculas (DUVENAUD et al., 2015).

2.3.1 *Spatial-Temportal Graph Convolutional Neural Network*

A *Spatial-Temporal Graph Convolutional Neural Network* (ST-GCN) é um modelo de RNA que utiliza os conceitos de RNC (Seção 2.2), RNR (Seção 2.1) e GNC (Seção 2.3) para analisar dependências temporais e estruturais simultaneamente em problemas que podem ser representados no formato de grafos (WU et al., 2019). Por exemplo, para reconhecer uma ação sendo realizada em uma sequência de esqueletos 3D, deve-se realizar a análise estrutural de cada esqueleto para obter o gesto representado, além de fazer a análise temporal entre esses gestos para classificar a ação sendo realizada. A Figura 13 ilustra esse processo.

Figura 13 – Exemplo da operação realizada por uma ST-GCN para classificar ações em sequências de esqueletos 3D.

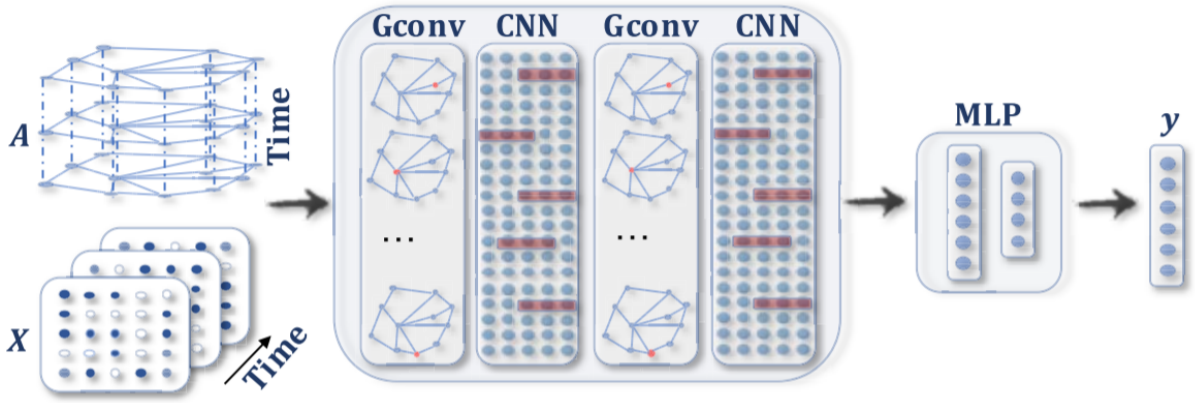


Fonte: (YAN; XIONG; LIN, 2018)

Existem duas abordagens para a análise temporal realizada pela ST-GCN, onde uma envolve RNR e outra RNC. Uma vez que as abordagens baseadas em RNR sofrem problemas baseados em desempenho computacional (WU et al., 2019), este trabalho irá focar nos modelos baseados em RNCs. A Figura 14 ilustra uma ST-GCN utilizada para a predição de tráfego em rodovias em Yu, Yin e Zhu (2018). Neste caso, os nós dos grafos representam sensores de velocidades e as arestas representam as distancias entre esses sensores. Assim, as dependências estruturais são obtidas por convoluções em grafos realizadas sobre A e as dependências temporais pela aplicação de uma RNC 1D em X. Por último, a saída do modelo é consumida por uma RNA MLP que gera a predição para o estado futuro do grafo.

Neste trabalho, modelos baseados na ST-GCN proposta em Yan, Xiong e Lin (2018) serão utilizados como um reconhecedores de ações humanas sendo realizadas em uma sequência de esqueletos 3D e são descritos nas Seções 2.3.2 e 2.3.3. As Figuras 26 e 13 também ilustram como os grafos de entrada serão criados, onde os nós são as juntas 3D que se encontram conectadas de acordo com a estrutura do corpo humano. Além disso, há arestas que conectam cada

Figura 14 – ST-GCN utilizada para a predição de tráfego em (YU; YIN; ZHU, 2018).



Fonte: (WU et al., 2019)

junta 3D com sua respectiva representação no próximo *frame* para representar suas relações temporais.

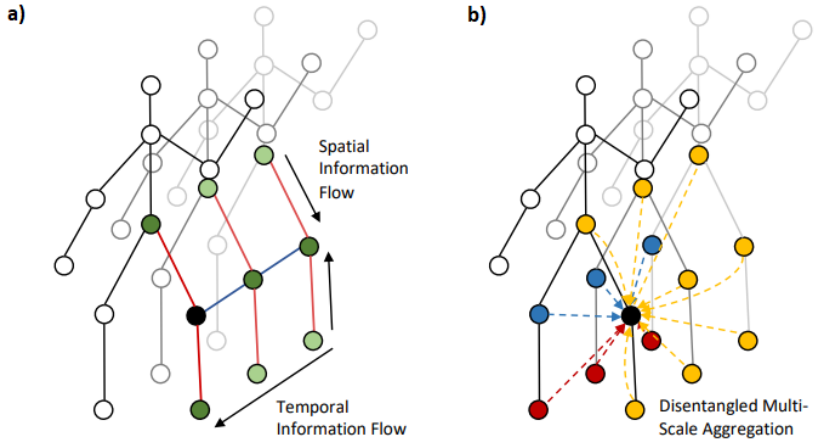
2.3.2 MS-G3D

O MS-G3D é um método de reconhecimento de ações, proposto em Ziyu Liu et al. (2020), que toma como base a modelagem espacial e temporal apresentada pela ST-GCN em Yan, Xiong e Lin (2018). O método busca tratar de dois problemas que ocorrem na ST-GNC: a representação do esqueleto 3D causa dificuldades para o reconhecedor aprender relações de juntas distantes em múltiplas escalas no fluxo temporal e a dificuldade de captura de relações complexas. Assim, esses problemas são tratados através da adição de pesos para as vizinhas de juntas tanto no fluxo temporal quanto no espacial.

A Figura 15 mostra a representação de um esqueleto 3D padrão na esquerda, onde o processamento ocorre de forma espacial e temporal separadamente. Por outro lado, o MS-G3D implementa a representação ilustrada na direita desta Figura, onde novas conexão são criadas para os nós no fluxo espacial e suas vizinhanças são segregadas de acordo com suas distâncias, sendo elas destacadas por cores diferentes neste caso.

Para atingir uma ponderação das vizinhanças das juntas em uma distância de k passos, deve-se utilizar uma matriz de adjacência A com uma ordem polinomial k, porém essa representação gera um viés nos nós mais próximos ao nó central, uma vez que a estrutura de grafo utilizada cíclica. Esse problema é tratado por meio da aplicação da Equação 9, que busca remover relações entre nós redundantes, onde $\tilde{A} = A + I$ e $d(v_i,v_j)$ é a distância mais curta entre v_i

Figura 15 – Representação de um esqueleto 3D: (a) é a representação padrão utilizada por (YAN; XIONG; LIN, 2018) e (b) é a representação proposta pelo MS-G3D em (LIU, Z. et al., 2020)



Fonte: (LIU, Z. et al., 2020)

e v_j.

$$[\tilde{A}_{(k)}]_{i,j} = \begin{cases} 1 \text{ se } d(v_i,v_j) = k \\ 1 \text{ se } i = j \\ 0 \text{ caso contrário} \end{cases} \tag{9}$$

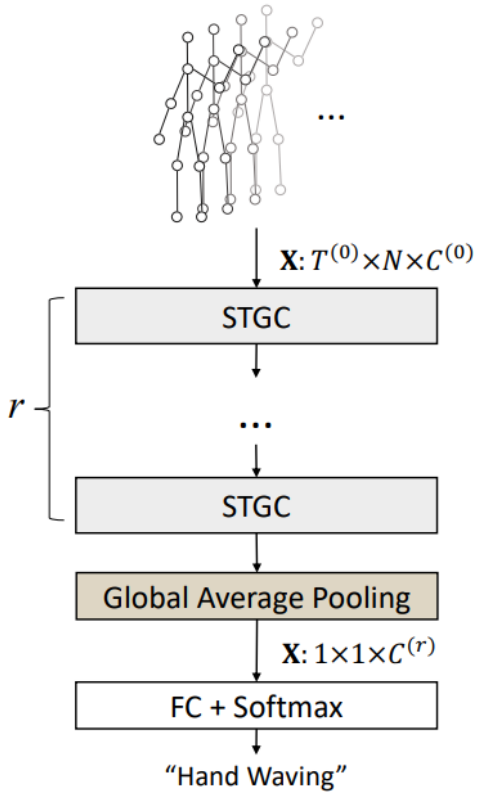
Assim, $\tilde{A}(k)$ se torna a representação da vizinhança em k passos de distância do nó central e a equação da convolução em grafos (Equação 8) é adaptada para a Equação 10, onde $H_t^{(l+1)}$ é a matriz de ativações para a camada l no tempo t e, para obter-se as relações entre juntas distantes, a matriz de ativação deve ser gerada para diversos valores de k.

$$H_t^{(l+1)} = \sigma(\sum_{k=0}^{K} \tilde{D}_{(k)}^{-\frac{1}{2}} \tilde{A}_{(k)} \tilde{D}_{(k)}^{-\frac{1}{2}} H_t^{(l)} W_{(k)}^{(l)}) \tag{10}$$

Por último, para melhorar o desempenho do reconhecedor na aprendizagem de relações espaciais e temporais complexas, a convolução em grafo é adaptada para a Equação 11, onde τ é a quantidade de *frames* em uma *sliding window* temporal, $\tilde{D}_{(\tau,k)}$ e $\tilde{A}_{(\tau,k)}$ tem uma definição semelhante aos seus valores anteriores, porém para os τ *frames* selecionados. Esta alteração tem como objetivo possibilitar que as convoluções sejam realizadas de forma espacial e temporal em conjunto, ao contrário do modelo anterior que as realiza de forma separada. Assim, é possível remover informações espaciais redundantes e permite o aprendizado de características temporais mais complexas.

$$[H_{(\tau)}^{(l+1)}]_t = \sigma(\sum_{k=0}^{K} \tilde{D}_{(\tau,k)}^{-\frac{1}{2}} \tilde{A}_{(\tau,k)} \tilde{D}_{(\tau,k)}^{-\frac{1}{2}} [H_{(\tau)}^{(l)}]_t W_{(k)}^{(l)}) \tag{11}$$

Figura 16 – Arquitetura do reconhecedor de ações MS-G3D.



Fonte: (LIU, Z. et al., 2020)

A Figura 16 ilustra a arquitetura do MS-G3D, que demonstrou desempenhos estado da arte em bases de dados volumosas disponibilizadas publicamente. Assim, devido ao seu alto desempenho, este reconhecedor será utilizado no trabalho proposto para reconhecer as ações em duas bases de dados selecionadas.

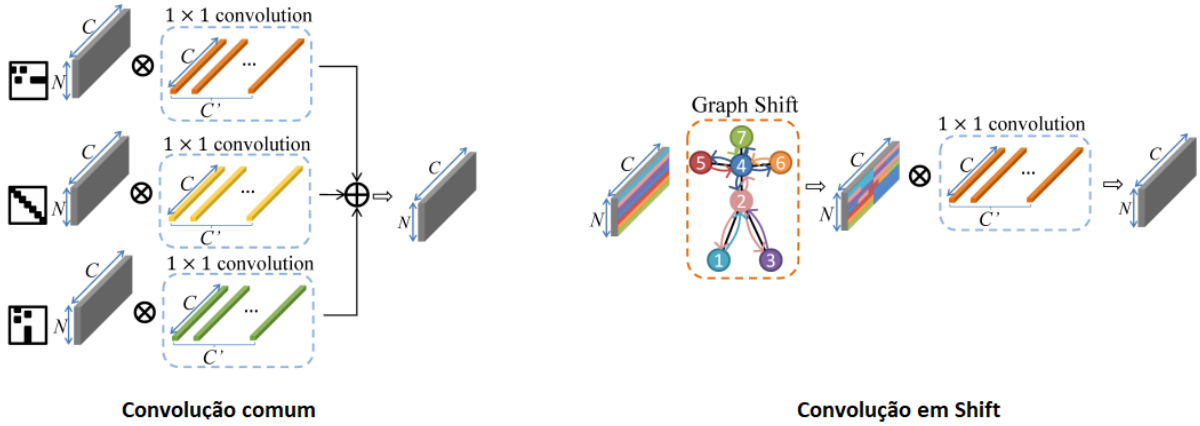
2.3.3 Shift-GCN

O Shift-GCN é um reconhecedor proposto em Cheng et al. (2020) com o objetivo de melhorar o desempenho computacional das convoluções em grafos e flexibilizar os campos receptivos da representação do esqueleto 3D tanto para sua parte espacial quanto temporal. Esses objetivos são atingidos através da adaptação das convoluções em grafo para operações de convoluções em *shift* ponto a ponto, que serão detalhadas nos próximos parágrafos.

Assim como é o caso das convoluções em RNCs, onde um *kernel* é aplicado nos dados de entrada e descolado em um passo s, o Shift-GCN cria um mecanismo o qual as convoluções do grafo também ocorrem por um *kernel* que se desloca pelo seu *input* realizando seu processa-

mento ponto a ponto de acordo em um determinado passo. A Figura 17 ilustra a diferença entre uma operação de convolução em grafo comum, na esquerda, e a operação proposta pelo trabalho, na direita. A operação comum realiza o processamento nos canais de seu *input* em *kernels* fixos para cada canal, enquanto a convolução em *shift* se desloca realizando o processamento ponto a ponto.

Figura 17 – Comparação entre a convolução em grafo comum (esquerda) e convolução em grafo em *shift* (direita).



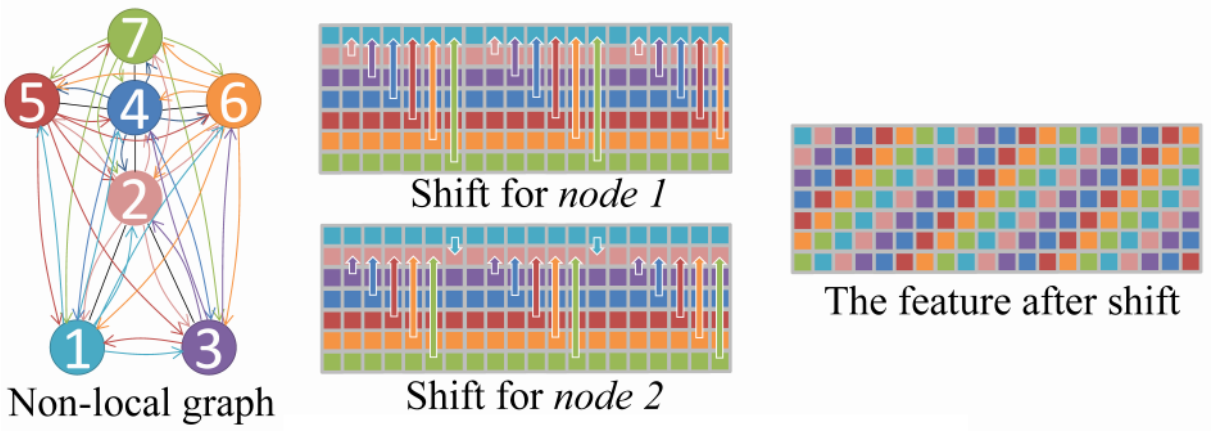
Fonte: (CHENG et al., 2020)

Para a análise espacial, uma convolução em *shift* não local é proposta. Assim, além da operação de deslocamento do *kernel* de convolução, cada convolução leva em consideração todos os nós do esqueleto 3D e não somente aqueles que são adjacentes. Essa operação se torna importante para representar certas relações, uma vez que ações tais como "bater palmas"ou "ler um livro"envolvem ambas as mãos do corpo humano, apesar delas serem representadas por nós distantes. A Figura 18 exemplifica este procedimento e é possível observar que, para cada nó, as informações de todos outros nós do grafo são recebidas.

Por último, para a convolução temporal, busca-se flexibilizar seu campo receptivo, uma vez que diversas camadas podem necessitar de campos receptivos diferentes. Para cada canal, um parâmetro S_i sendo $i = 1,2,...,C$, é aplicado para o calculo do passo de deslocamento do *kernel*. Por outro lado, o passo é calculado por uma interpolação linear e, uma vez que essa função é diferenciável, S_i pode ser aprendido durante o treinamento do reconhecedor.

O Shift-GCN se mostrou eficiente computacionalmente e demonstrou desempenho estado da arte em bases de dados volumosas, logo este reconhecedor será utilizado no trabalho proposto para reconhecer as ações nas duas bases de dados selecionadas.

Figura 18 – Exemplo de uma convolução em grafos em *shift*.



Fonte: (CHENG et al., 2020)

2.4 VALIDAÇÃO CRUZADA

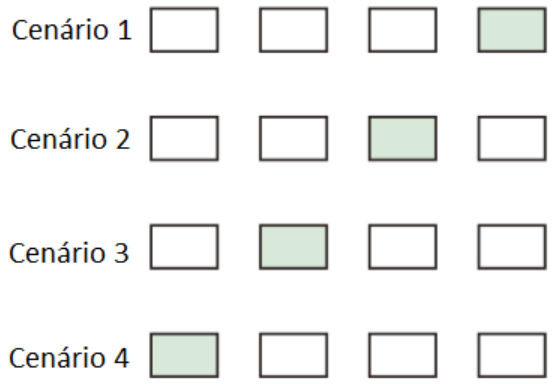
A validação cruzada é um método estatístico utilizado para o treinamento e validação de algoritmos de aprendizado de máquina. Assim, dado um conjunto N de modelos candidatos, deseja-se comparar esses modelos e decidir qual deles apresenta um melhor desempenho para o problema em questão. Após a escolha do modelo, uma última validação é realizada para verificar se o modelo escolhido não representou um melhor desempenho devido a um possível *overfitting* nos dados de treinamento (HAYKIN, 2009).

Assim, para executar a validação cruzada, deve-se particionar uma base de dados em dois conjuntos, sendo eles o conjunto de treinamento e conjunto de teste, que são utilizados para escolher o melhor modelo e para validá-lo, respectivamente. Além disso, o conjunto de treinamento deve ser repartido em dois subconjuntos, onde um será utilizado para treinar os modelos candidatos e outro para verificar qual é o de melhor desempenho. Por último, o modelo escolhido deverá ser validado por meio do conjunto de testes, onde será possível verificar se houve especialização do modelo no conjunto de treinamento.

Além disso, há diversas variações do método de validação cruzada tais como a validação cruzada *multifold* e o método *leave-one-out*. O método de validação cruzada *multifold* é geralmente utilizado em cenários onde a quantidade de amostras de treino são reduzidas. Assim, uma base de dados contendo N amostras pode ser particionada em K partições, sendo $K > 1$, onde o modelo é treinado em $K - 1$ partições e a partição restante é utilizada para a validação do modelo (HAYKIN, 2009). Por último, o desempenho do modelo poderá ser obtido pela

aplicação do método para todas as possibilidades de iterações entre partições de treinamento e validação. A Figura 19 ilustra esse procedimento.

Figura 19 – Exemplo da validação para o método *multifold* com $K = 4$, onde todas as interações entre os $K - 1$ conjuntos de treinamento (em branco) e o conjunto validação (em verde) são testadas.



Fonte: (HAYKIN, 2009).

Por outro lado, quando a quantidade de amostras disponíveis N for muito reduzida, o método *leave-one-out* poderá ser utilizado. Assim, este método busca treinar o modelo com $N - 1$ amostras e validá-lo com a amostra restante, onde esse procedimento é realizado para as N possibilidades de partição entre as amostras de treinamento e amostra de validação.

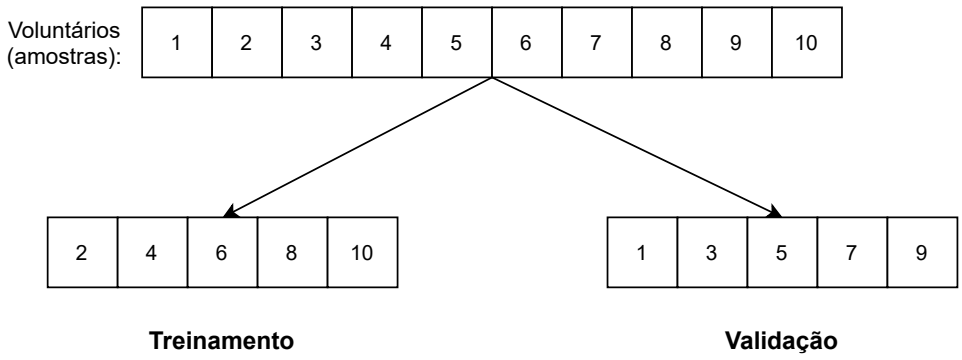
Na área de reconhecimento de ações humanas, é possível observar a utilização de 4 variações da validação cruzada: *cross-subject*, *cross-view*, *cross-setup* e *leave-one-out*. A validação *cross-subject* (Seção 2.4.1) envolve a partição da base de dados pela quantidade de voluntários e é utilizada para verificar a tolerância do método de reconhecimento a variação das ações realizadas por pessoas diferentes. A validação *cross-view* (Seção 2.4.2) particiona a base de dados em relação às perspectivas de captura de vídeo para verificar a tolerância do reconhecedor a alterações de perspectiva. A validação *cross-setup* (Seção 2.4.3) tem como objetivo particionar a base de acordo com ambientes e perspetivas de coletas diferentes. Por último, como há poucas bases de dados volumosas para essa tarefa, a validação *leave-one-out* é amplamente utilizada para bases de dados pequenas e envolve a utilização das amostras de 1 voluntário como validação do método de reconhecimento.

2.4.1 Validação cross-subject

Um dos desafios do reconhecimento de ações humanas é manter um bom desempenho apesar das variações de movimentos que podem ser observadas entre pessoas. Por exemplo, uma pessoa idosa faz movimentos diferentes e com durações diferentes para levantar de uma cadeira quando comparada com uma pessoa jovem. Assim, a validação *cross-subject* busca avaliar a robustez de um classificador em relação a essas diferenças e é utilizada em trabalhos como Shahroudy et al. (2016) e Jun Liu et al. (2019).

O método envolve o particionamento de uma base de dados em treinamento e validação, onde amostras de um mesmo usuário não podem estar presentes nas duas partições ao mesmo tempo. A Figura 20 ilustra um exemplo de aplicação da validação *cross-subject* para um conjunto de 10 voluntários. Neste caso, os voluntários com o número de identificação pares compõem a base de treinamento, enquanto os com identificação ímpar compõem a base de validação.

Figura 20 – Exemplo de partição realizada pelo método *cross-subject*.



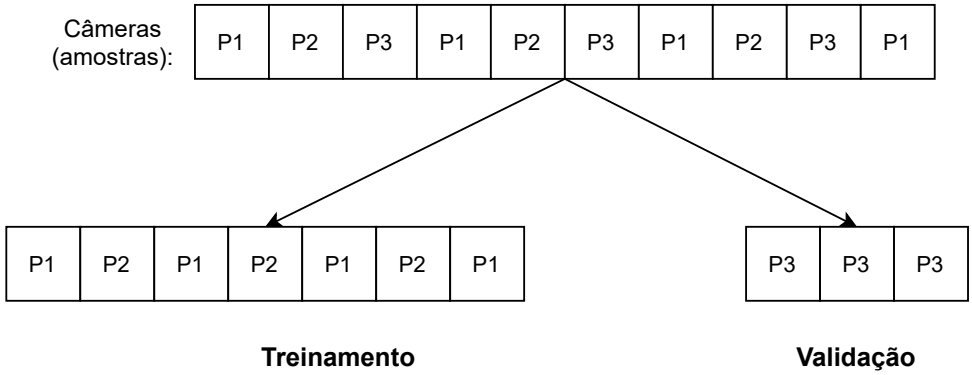
Fonte: Autor.

2.4.2 Validação cross-view

Além das variações causadas por voluntários diferentes, também é possível observar variações causadas por perspectivas diferentes. Por exemplo, um classificador de ações treinado em uma base de dados com vídeos frontais pode apresentar dificuldades para reconhecer as mesmas ações em vídeos laterais. Assim, a validação *cross-view* busca avaliar o desempenho de um classificador em perspectivas diferentes e é utilizado em trabalhos como Shahroudy et al. (2016) e Yan, Xiong e Lin (2018).

O método envolve o particionamento de uma base de dados em treinamento e validação, onde amostras de uma perspectiva não podem pertencer aos dois conjuntos ao mesmo tempo. A Figura 21 mostra um exemplo, onde a base de dados contem as perspectivas $P1$, $P2$ e $P3$. Neste caso, a base de dados de treinamento é composta por $P1$ e $P2$ e a base de validação é composta por $P3$.

Figura 21 – Exemplo de partição realizada pelo método *cross-view*.



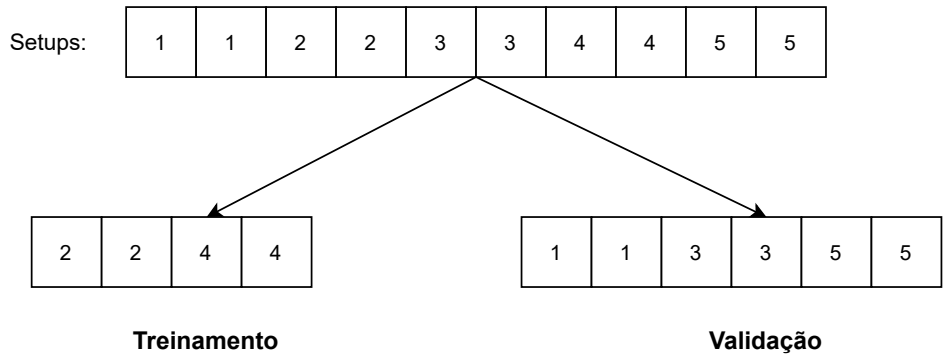
Fonte: Autor.

2.4.3 Validação cross-setup

De forma semelhante a validação *cross-view* que busca verificar a robustez dos reconhecedores em relação a perspectivas diferentes, a validação *cross-setup* também busca avaliar o desempenho dos reconhecedores em relação a *setups* diferentes, onde um *setup* é composto por um local e uma perspectiva de coleta diferentes e pode ser observada em trabalhos como Jun Liu et al. (2019) e Ziyu Liu et al. (2020).

O método tem como objetivo particionar a base de dados em um conjunto de treinamento e outro de validação, onde amostras amostras de um conjunto de *setups* são utilizadas para o treinamento e as restantes para a validação. A Figura 22 mostra um exemplo, onde a base de dados é composta por 5 *setups* diferentes e as amostras dos *setups* de número par são utilizados para treinamento e as com *setups* de número ímpar utilizadas para validação.

Figura 22 – Exemplo de partição realizada pelo método *cross-setup*.



Fonte: Autor.

2.5 RECONHECIMENTO DE AÇÕES EM VÍDEOS RGB+D

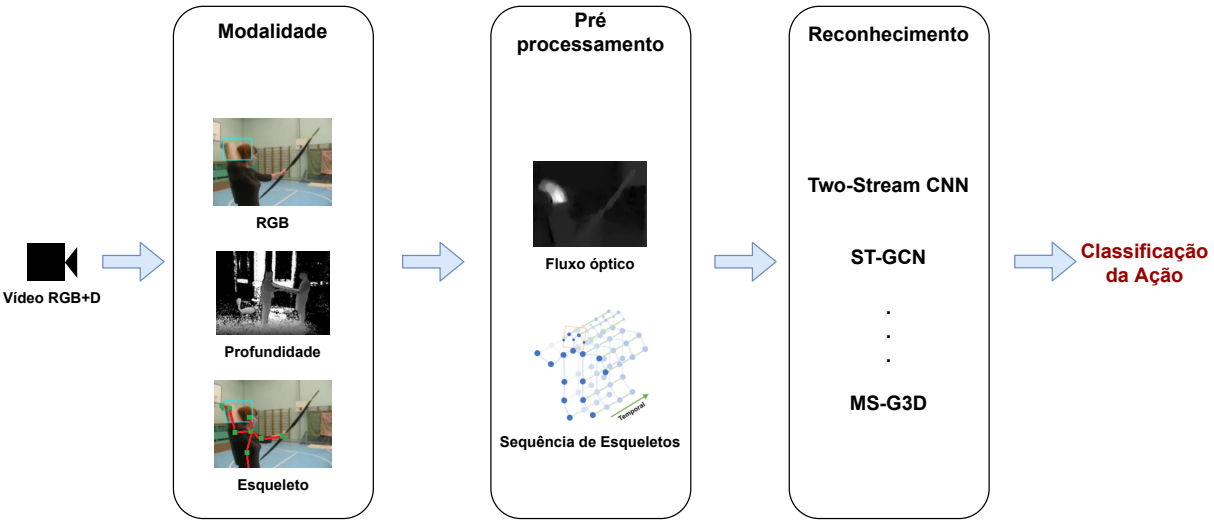
O problema de reconhecimento de ações em vídeos é uma tarefa desafiadora que vêm despertando o interesse de pesquisadores devido as suas diversas possibilidades de aplicação. Por outro lado, ao se comparar com o reconhecimento de imagens, o reconhecimento de ações apresenta uma maior complexidade devido a análise de relações temporais entre os *frames* capturados. Além disso, as variações de movimentações entre indivíduos e perspectivas diferentes adicionam mais um desafio no processo de reconhecimento, tornando-se necessária uma quantidade volumosa de dados para o treinamento dos reconhecedores (FEICHTENHOFER; PINZ; ZISSERMAN, 2016).

Para realizar o reconhecimento de ações humanas em vídeos RGB+D, um método deve seguir os seguintes passos: seleção da modalidade de dados (Seção 2.5.1), pré-processamento das amostras (Seção 2.5.2) e reconhecimento (Seção 2.5.3). A Figura 23 ilustra este processo, onde as próximas Seções detalham as tarefas que o compõem e os possíveis desafios que podem ser encontrados.

2.5.1 MODALIDADE DE DADOS

O primeiro passo do reconhecimento de ações é definir a modalidade de dados a qual será analisada. As principais modalidades de dados na literatura são *frames* RGB, *frames* de profundidade e esqueletos 3D. A Figura 24 ilustra amostras de cada modalidade onde o *frame* RGB, na esquerda, representa os pixeis de cores de uma imagem, o *frame* de profundidade, no

Figura 23 – Processo de reconhecimento de ações em vídeos RGB+D.



Fonte: Autor.

centro, representa o mapa de profundidade da imagem e o esqueleto, na direita, representa a pose do corpo humano no *frame*.

Figura 24 – Modalidades de dados para o reconhecimento de ações.

Fonte: (SIMONYAN; ZISSERMAN, 2014).

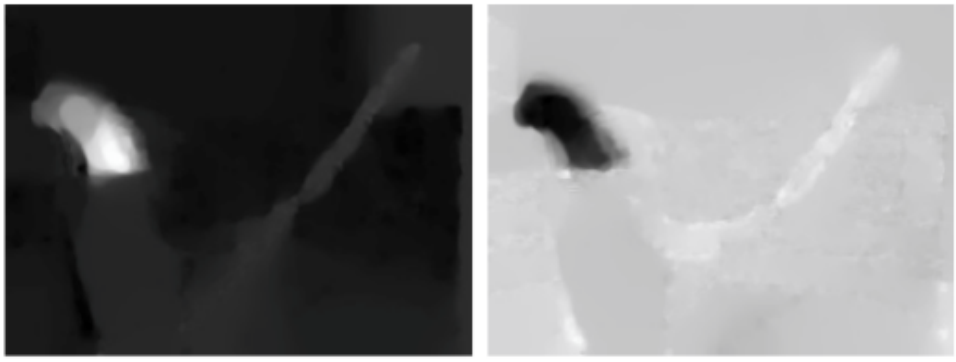
A escolha da modalidade de dados ou métodos de reconhecimento é realizada de acordo com os recursos disponíveis. Se existir modalidades de dados limitadas, deve-se escolher reconhecedores que trabalham com as mesmas. Por outro lado, se for necessária a execução de um reconhecedor específico, os dados inseridos terão que seguir seus requisitos. É importante avaliar as informações que cada modalidade pode fornecer. Por exemplo, em Simonyan e Zisserman (2014) e Feichtenhofer, Pinz e Zisserman (2016) , *frames* RGB são utilizados para obter informações espaciais, como objetos e poses, e informações temporais como o fluxo de movimentação observado durante uma sequência de *frames*. Por outro lado, trabalhos como Yan,

Xiong e Lin (2018) utilizam grafos compostos por esqueletos 2D ou 3D para representar as movimentações realizadas por um indivíduo.

2.5.2 PRÉ-PROCESSAMENTO DAS AMOSTRAS

Uma vez que a modalidade foi escolhida, suas amostras devem ser pré-processadas para serem alimentadas ao reconhecedor. Por exemplo, dada uma sequência de *frames* RGB, pode-se calcular o fluxo óptico entre os mesmos assim como é ilustrado pela Figura 25, onde as componentes verticais desse fluxo são representadas pela Figura 25 (a) e as componentes verticais pela Figura 25 (b). Além disso, deve-se levar em consideração limitações de hardware uma vez que a representação da amostra obtida após o pré-processamento pode ter um alto custo computacional ou até mesmo ocupar um grande volume de armazenagem.

Figura 25 – *Frames* de profundidade em dois componentes: (a) imagem da componente horizontal e (b) imagem da componente vertical.



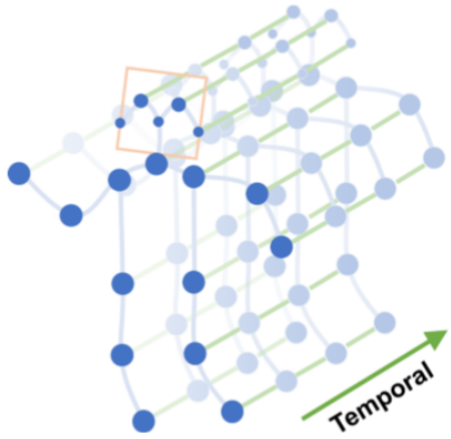
(a) (b)

Fonte: (SIMONYAN; ZISSERMAN, 2014)

Em trabalhos como Feichtenhofer, Pinz e Zisserman (2016), os fluxos ópticos utilizados foram obtidos através do algoritmo TVL1, que é um algoritmo computacionalmente custoso. De acordo com Medhekar et al. (2019), para obter o fluxo óptico denso TVL1 entre dois *frames* RGB, uma placa de vídeo com arquitetura Turing utiliza cerca de 400 milisegundos de tempo computacional. Assim, caso a base de dados seja muito volumosa, um tempo de processamento alto será necessário. Além disso, para otimizar o volume de armazenamento, cada canal das amostras obtidas foi armazenado como uma imagem JPEG em virtude da compressão fornecida pelo formato, portanto, como cada amostra gerada contem 20 canais, 20 imagens JPEG são obtidas.

A modalidade de esqueletos 2D ou 3D é utilizada por trabalhos recentes na literatura tais como Yan, Xiong e Lin (2018) e Ziyu Liu et al. (2020), pois suas amostras podem ser processadas e transformadas em grafos representando as movimentações realizadas por um indivíduo (Figura 26) com tamanho e tempo computacional menor do que o cálculo de fluxos ópticos, além de ser não apresentar ruídos ambientais tais como o variações de luz, fundo do *frame* e roupas utilizadas pelo indivíduo.

Figura 26 – Amostra gerada a partir de uma sequência esqueletos 2D ou 3D.



Fonte: (YAN; XIONG; LIN, 2018)

Por último, é possível observar um aumento de desempenho de métodos que empregam mecanismos de atenção para a seleção dos *frames* e informações utilizadas. Isso pode ser verificado em trabalhos como Song et al. (2018) e Dong Li et al. (2019), onde a seleção prévia dos *frames* resultou no aumento de desempenho do reconhecedor. Neste caso, esses mecanismos são aplicados antes do processamento e geração da amostra que será alimentada ao reconhecedor.

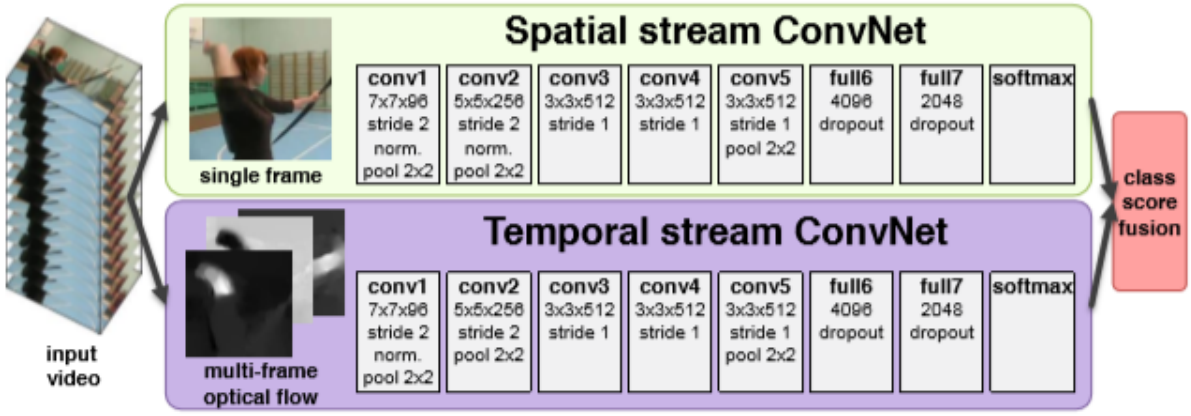
2.5.3 RECONHECIMENTO

A última etapa do reconhecimento de ações em vídeos é a análise das amostras de dados através de um algoritmo de classificação, assim as amostras são alimentadas e seus rótulos são obtidos. Existem diversas abordagens para realizar essa etapa, porém é possível observar a crescente utilização de métodos de *deep learning* na literatura recente, sendo eles baseados em RNCs, GNCs e RNRs. Por exemplo, Ji et al. (2013) e Yan, Xiong e Lin (2018) utilizam

métodos baseados em RNC e GNC, respectivamente, e apresentaram desempenho estado da arte nas bases de dados mais volumosas da literatura.

Além disso, para capturar uma maior quantidade de informações da amostra, existem reconhecedores que utilizam mais de uma modalidade de fluxo de processamento. A Figura 27 representa o reconhecedor utilizado em Simonyan e Zisserman (2014) onde o reconhecimento de ações é realizado por 2 fluxos: uma para o processamento espacial utilizando *frames* RGB e outro para a analise temporal utilizando o fluxos óptico de 10 *frames* consecutivos. Apesar dessa abordagem resultar em um maior desempenho do reconhecedor, os modelos se tornam mais complexos, logo uma base de dados mais volumosa e técnicas de regularização devem ser consideradas devido a facilidade de *overfittig*. Por último, caso uma base de dados volumosa não esteja disponível, métodos de *fine-tunning* e transferência de aprendizagem podem ser utilizados ou reconhecedores que utilizam descritores manuais tais como Dobhal et al. (2015) e Chen, Liu e Kehtarnavaz (2013) podem ser aplicados.

Figura 27 – Exemplo de um reconhecedor com dois fluxos de processamento.



Fonte: (SIMONYAN; ZISSERMAN, 2014)

3 REVISÃO BIBLIOGRÁFICA

Uma vez que o trabalho proposto busca explorar o reconhecimento de comportamento humano no nível semântico de ações, esta Seção tem como objetivo detalhar a área da bibliografia estudada para o seu desenvolvimento. Assim, a Seção 3.1 descreve as bases de dados mais utilizadas e a Seção 3.2 detalha os métodos de reconhecimento de ações apresentados na literatura recente.

3.1 BASES DE DADOS RGB+D

As bases de dados para o reconhecimento de ações podem apresentar diversas características que as distinguem umas das outras. Por exemplo, algumas bases podem ser pouco volumosas e ter como objetivo o reconhecimento de um grupo pequeno de ações. Por outro lado, existem bases volumosas coletadas para o treinamento de algoritmos que necessitam de um grande de volume de dados para seu treinamento. Também existem bases coletadas de vídeos em plataformas tais como Youtube ou de filmes e que disponibilizam somente o rótulo da amostra, enquanto existem outras bases que também contêm informações sobre o voluntário, perspectiva e *setup* de coleta. Assim, os próximos parágrafos buscam descrever a trajetória e as características das bases de dados de reconhecimento de ações mais utilizadas na literatura estudada.

Uma base e de dados para o reconhecimento de ações denominada Weizmann é proposta em Blank et al. (2005). A base é composta por 81 amostras que foram capturadas por 9 voluntários onde cada um realizou 9 ações diferentes em um fundo estático. A base foi testada por um reconhecedor pelo método *leave-one-out*, onde uma taxa de erro de classificação de 6,38% foi obtida.

Em Li, Zhang e Liu (2010), um método de reconhecimento de ações e uma base de dados, denominada MSRAction3D, baseados em mapas de profundidade. A base de dados foi coletada por 10 voluntários e contem 567 amostras dividas em 20 ações e uma única perspectiva. Além disso, o reconhecedor proposto processa mapas de profundidade para extrair esqueletos 3D que são consumidos por um *action graph* para sua classificação. Experimentos foram realizados na base de dados proposta pelos métodos da validação cruzada e *cross-subject* e os desempenhos de 94,2% e 74,7% foram obtidos.

Uma base de dados contendo 6766 amostras de vídeos retirados de filmes ou da internet, denominada HMDB51, é proposta em Kuehne et al. (2011). Os vídeos foram coletados por meio da segmentação de trechos de filmes e vídeos disponíveis online. Os seguimentos que compõem a base de dados foram obtidos de 101 segmentos selecionados, sendo eles divididos em 51 classes. Experimentos foram realizados na base de dados com dois algoritmos de reconhecimento de ações, onde ambos mostraram um desempenho próximo a 20%. Um experimento com um subconjunto da base também foi realizado onde um desempenho de 54,3% foi obtido.

Em J. Wang et al. (2012), base de dados denominada MSRDailyActivity3D é proposta. A base de dados coletada contem 320 amostras dividas em 16 classes, onde cada voluntário as executou 2 vezes em uma única perspectiva. Um novo reconhecedor de ações baseado em *Action Esemble* foi proposto para validar a base pelo método *cross-subject*, onde obteve uma precisão de 85,75% foi obtida, respectivamente.

A base de dados UCF101 composta por 13320 amostras coletadas da plataforma de vídeos Youtube é proposta em Soomro, Zamir e Shah (2012). As amostras foram coletadas de vídeos do Youtube, onde os mesmos foram rotulados em 101 classes de ações diferentes. Um reconhecedor de ações baseado em descritores manuais e SVM foi utilizado para validar a base de dados pelo método da validação cruzada e um desempenho de 44,5% foi obtido.

Uma base de dados, denominada UTKinect-Action, contendo 200 amostras é proposta em Xia, Chen e Aggarwal (2012). As amostras foram coletadas por meio de um sensor Kinect e as modalidades de *frames* RBG, *frames* de profundidade e esqueletos 3D foram disponibilizadas. Além disso, a base de dados foi coletada por 10 voluntários onde cada um realizou 10 tipos de ações 2 vezes. Um novo reconhecedor denominado *Histograms of 3D Joint Locations* (HOJ3D) foi proposto para a validação da base, onde um desempenho de 90,92% foi obtido pelo método *leave-one-out*.

Em Ofli et al. (2013), uma base de dados denominada *Multimodal Homam Action Database* (MHAD) composta 660 amostras é proposta. Os vídeos se encontram separados em 11 classes de ações e foram coletados por 12 voluntários, sendo que cada ação foi gravada 5 vezes para cada voluntário nas modalidades de *frames* RGB, *frames* de profundidade e áudio. Um algoritmo multimodal de reconhecimento de ações foi aplicado na base de dados para a validação pelo método *cross-subject*, onde 92,27% de desempenho foi obtido para a junção de informações de *frames* RGB e de profundidade.

A base de dados DAily Home LIfe Activity (DAHLIA), com foco no monitoramento de idosos e aplicações de *smart home*, é proposta em Vaquette et al. (2017). Ela foi coletada por 44 voluntários realizando 7 atividades de longo termo com um média de 39 minutos de duração, totalizando 51 amostras. A base de dados foi comparada com outras bases da literatura onde foi possível observar uma maior duração média de suas amostras e maior complexidade das atividades coletadas.

Uma base de dados volumosa contendo 56880 amostras coletadas em 3 modalidades diferentes denominada NTU RGB+D é proposta em Shahroudy et al. (2016). A base de dados contem 60 classes de ações coletadas por 40 voluntários em 80 perspectivas diferentes, além de disponibilizar as modalidades de *frames* RGB, *frames* de profundidade e esqueletos 3D. Para validar a base de dados, um reconhecedor denominado P-LSTM foi criado, onde um sequência de esqueletos 3D é consumida para classificar a ação sendo realizada. O reconhecedor foi validado e comparados com outros métodos estado da arte pela validação *cross-subject* e *cross-view*, onde obteve os maiores desempenhos de 62,93% e 70,27%, respectivamente.

Em Jun Liu et al. (2019), a base recebeu uma expansão que foi denominada NTU RGB+D 120. A nova base é composta por 14480 amostras coletadas por 106 voluntários em 155 perspectivas diferentes coletadas em 32 *setups*, além de disponibilizar as modalidades de *frames* RGB, *frames* de profundidade, esqueletos 3D e infravermelho. Reconhecedores estado da arte foram aplicados na base de dados onde desempenhos de até 64,6% e 66,9% foram obtidos para os métodos do *cross-subject* e *cross-setup*.

Em Kay et al. (2017), uma base de dados volumosa contendo 306245 amostras denominada Kinetics é proposta. Os vídeos foram coletados do Youtube por meio de uma ferramenta de busca e revisão de voluntários. A modalidade de dados disponibilizada é *frames* RGB e suas amostras se encontram dividas em 400 classes. Reconhecedores estado da arte foram aplicados na base de dados pelo método da validação cruzada onde desempenhos de até 61% foram obtidos. A base sofreu diversas expansões sendo, a mais recente em Smaira et al. (2020), resultando em uma base composta por cerca de 650000 vídeos classificados em 700 ações diferentes.

A Tabela 1 resume as características de cada uma dessas bases de dados e é possível observar que elas encontram dividas em duas metodologias principais de coleta de amostras: coleta via plataformas de conteúdo e coleta por meio de voluntários. No caso da coleta por plataformas, os vídeos de amostras são coletados por meio de plataformas de vídeos e filmes online como o Youtube, sendo as bases de dados UFC101 e HMDB51 exemplos deste grupo. Uma das vantagens deste modelo é que a coleta de amostras pode ser feita de forma rápida

Tabela 1 – Comparação das Bases de Dados. As modalidades de dados são: RGB para imagens RGB, D para mapas de profundidade, 3DJ para juntas 3D e IR para infravermelho.

Base de Dados Ano	Vídeos	Classes	Voluntários	Perspectivas	Modalidade
Weizmann 2005	81	9	9	1	RGB
MSRAction3D 2010	567	20	10	-	D
HMDB51 2011	6766	51	-	-	RGB
MSRDailyActivity3D 2012	320	16	-	-	RGB+D
UCF101 2012	13320	101	-	-	RGB
UTKinect-Action 2012	200	10	10	-	RGB+D+3DJ
MHAD 2013	660	11	12	-	RGB+D
NTU RGB+D 2016	56880	60	40	80	RGB+D+3DJ+IR
DAHLIA 2017	51	7	44	-	RGB+D+3DJ
Kinetics400 2017	306245	400	-	-	RGB
NTU RGB+D 120 2019	114480	120	106	155	RGB+D+3DJ+IR
Kinetics-700-2020 2020	647907	700	-	-	RGB

Fonte: Autor.

e com o auxílio de ferramentas de busca, assim é possível verificar que as bases de dados que utilizaram esta metodologia contêm um volume de dados maior em comparação com as outras bases propostas em datas próximas. Por exemplo, A UCF101 foi proposta em 2012 e disponibiliza 13320 amostras, enquanto a base de dados MHAD, proposta em 2013, contem somente 660 amostras. Em contrapartida, enquanto as bases onde a coleta foi realizada por meio de voluntários geralmente possuem menos amostras, suas anotações são mais ricas uma vez que possuem informações úteis para as validações dos reconhecedores tais como os voluntários responsáveis por cada amostra, perspectiva de coleta e *setup* de gravação. Logo, é possível verificar a robustez do reconhecedor em relação a diferentes voluntários, perspectivas e *setups* de gravação.

Uma vez que neste trabalho tem como um dos objetivos avaliar os reconhecedores em relação a diferentes voluntários e perspectivas em diferentes cômodos de um ambiente doméstico, as bases que serão consideradas para o treinamento dos reconhecedores serão aquelas coletadas por meio de voluntários e *setups* controlados tais como a NTU RGB+D 120.

3.2 RECONHECIMENTO DE AÇÕES

Existem diversos métodos de reconhecimento de ações na literatura que utilizam diversos tipos de descritores e classificadores. Por exemplo, métodos mais antigos geralmente utilizam descritores manuais e classificadores como a SVM e, por outro lado, é possível observar uma crescente utilização de descritores e classificadores baseados em *deeplearning* em trabalhos recentes. Também é possível observar que existem diversas modalidades de dados que podem ser utilizadas pelos reconhecedores tais como *frames* RGB, *frames* de profundidade e esqueletos 3D. Esta Seção busca descrever outros trabalhos que exploram o reconhecimento de ações utilizando dados coletados por sensores RGB+D.

Uma técnica de reconhecimento de ações que utiliza informações de movimentações e objetos como forma de melhorar seu desempenho é proposta em Khan, Singh e Nevatia (2012). Os reconhecimentos de ações, pose e objetos são realizados de forma simultânea e seus descritores são armazenados em um modelo de grafo para cada *frame*. O método foi validado em uma base de dados composta por 6 tipos de ações coletada por 10 voluntários, onde 4 objetos estão envolvidos. O método de validação utilizado foi o *leave-one-out*, onde um desempenho de 92,31% foi obtido. Além disso, foi possível verificar que a adição das informações de pose e objeto resultam no aumento do desempenho do reconhecedor.

Em J. Wang et al. (2012), um método de reconhecimento de ações utilizando *actionlets* em mapas de profundidade e esqueletos 3D é proposto. O reconhecedor utiliza um descritor denominado *Local Occupancy Pattern* (LOP), que é calculado para cada junta de um esqueleto 3D. Além, disso, um *Action Esemble* é obtido por meio do cálculo de *actionlets* para diversos subconjuntos dessas juntas, que é então consumido por uma SVM durante os processos de treinamento e classificação. Experimentos foram realizados nas bases de dados CMU MoCap, MSRAction3D e MSRDailyActivity3D pelo método *cross-subject*, onde obteve uma precisão de 98,13% , 88,2% e 85,75%, respectivamente.

Um método de reconhecimento de ações que utiliza RNCs 3D para analisar imagens de câmeras de vigilância é proposto em Feichtenhofer, Pinz e Zisserman (2016). O vídeo de *input*

é consolidado em um bloco único, onde cada um de seus canais é consumido por uma RNC que gera descritores espaciais e temporais e, por último, o descritor final é obtido pela junção dos resultados de cada canal. Experimentos foram realizados nas bases de dados TRECVID 2008 e KTH por meio da validação cruzada e *cross-subject*, onde obteve desempenho de 88,53% e 90,2%. Trabalhos futuros envolvem utilizar um modelo pré-treinado com aprendizagem não supervisionada e o uso de *Deep Belief Neural Networks* para o reconhecimento de ações.

Em Wang e Schmid (2013), um reconhecedor de ações baseado em trajetórias densas e tolerante à movimentação de câmera é proposto. Os *frames* RGB de uma determinada sequência de *input* são comparados e a movimentação da câmera é obtida, através da comparação de seus descritores SIFT, e removida das trajetórias calculadas para cada *frame*. Por último, descritores são computados, normalizados e alimentados à uma SVM para a classificação. Experimentos foram realizados nas bases de dados Hollywood2, HMDB51, Olympic Sports e UCF50, onde obteve precisão de 64,3%, 57,2%, 91.1% e 91.2%, respectivamente. Os métodos de validação utilizados foram a validação cruzada para as bases Hollywood2 e Olympic Sports, a *cross-subject* para a base HMDB51 e a *leave-one-out* para a UCF50.

Um método de reconhecimento de ações que utiliza descritores *Depth Motion Maps* (DMM) e *Sparse Representation Classifier* (SRC) é proposto em Chen, Liu e Kehtarnavaz (2013). Os *frames* de profundidade de uma sequência são comparados com seus *frames* adjacentes, onde suas diferenças são obtidas e o DMM gerado. Assim, esse DMM é consumido por um SRC para treinamento e classificação. Experimentos foram realizados na base de dados MSRAction3D pelos métodos validação cruzada e *cross-subject* e os desempenhos de 99,1% e 96,2% foram obtidos.

Em Bingbing Ni et al. (2013), um método de detecção de ações utilizando Redes Bayesianas é proposto. Poses de interesse são obtidas através de *frames* de profundidade e *frames* RGB e, em seguida, esses dados são transformados em *tracklets* para consolidar as informações espaciais e temporais desses *frames*. Por último, descritores 3D são criados e uma Rede Bayesiana classifica a ação sendo realizada. Experimentos foram realizados na base de dados HARL 2012 e o método proposto foi comparado com outro método estado da arte e obteve um f-score de 0,346.

Em Kay et al. (2017), um método de reconhecimento de ações que utiliza dois fluxos de processamento, um espacial e outro temporal, é proposto. Os *frames* do vídeo de *input* são segregados em cortes e seus fluxos ópticos são calculados e consolidados em uma estrutura única. Assim, os *frames* RGB são alimentados a uma RNC que faz a análise espacial e os fluxos

ópticos a uma RNC que faz a análise temporal e, por último, seus resultados sofrem fusão para obter a classificação final da ação. Um modelo pré-treinado na base de dados ILSVRC-2012 foi utilizado para a validação pelo método *cross-subject* nas bases UCF-101 e HMDB-51 obtendo 88% e 59,4% de desempenho, respectivamente.

Uma junção de métodos baseados em RNCs e RNRs para o reconhecimento de ações em uma sequência de *frames* de tamanho variável é proposta em Donahue et al. (2014). O vídeo de entrada é alimentado a uma RNC para a extração de características espaciais, que são processadas por uma RNR para obter suas relações temporais. Experimentos foram realizados na base de dados TACoS, composta por 44762 amostras, onde o modelo foi validado pela métrica BLEU-4 com desempenho de 28.8%.

Em Foggia et al. (2014), um método de reconhecimento que utiliza *Deep Belief Networks* (DBN) em mapas de profundidade é proposto. Uma sequência de *frames* de profundidade é processada por uma DBN que utiliza *Restricted Boltzmann Machine* para obter descritores de alto nível. O método foi validado nas bases de dados MHAD e MIVIA pela validação cruzada e um desempenho de 84,7% foi obtido.

Um método de reconhecimento de ações baseado em RNC é proposto em Ijjina e Mohan (2014). O método utiliza técnicas de *motion capture* (MOCAP) de 3 juntas 3D de um individuo para criar descritores, que serão utilizados por uma RNC para a classificação de seu *input*. O método proposto foi validado na base de dados MHAD, onde obteve 98,38% de precisão pelo o método da validação cruzada.

Em Vemulapalli, Arrate e Chellappa (2014), um método de reconhecimento de ações baseado em relações geométricas entre juntas 3D é proposto. Informações de juntas de um esqueleto 3D são obtidas e suas relações geométricas são representadas por curvas, onde uma SVM é utiliza para sua classificação. Experimentos foram realizados nas bases de dados MSR-Action3D, UTKinect-Action e Florence3D-Action pelo método *cross-subject* e precisões de 92,46%, 97,08% e 90,88% foram obtidas, respectivamente.

Um método de reconhecimento de atividades em múltiplos pontos de vista, denominado FastDTW, é proposto em Agarwal, Kushwaha e Srivastava (2015). Os *frames* RGB de um vídeo são analisados e esqueletos 3D são extraídos e normalizados. Além disso, pesos são atribuídos às juntas de cada esqueleto de acordo com a atividade sendo realizada. Assim, após o processamento dos *frames* de entrada, o algoritmo DTW é aplicado para o reconhecimento da ação sendo realizada. Uma base de dados com 8 ações foi criada pelo autor do trabalho para a validação do método e uma precisão de 99% foi obtida.

Em Pichao Wang et al. (2015), um método de reconhecimento de ações, denominado *Hidden Depth Motion Maps* (HDMM), que realiza o reconhecimento de ações por meio de *frames* de profundidade é proposto. O método processa uma sequência de *frames* de profundidade e atribui pesos para cada um e, por último, os dados gerados são segmentados em 3 canais, um para cada eixo da amostra, e alimentados a uma RNC. O reconhecedor foi validado pela método *cross-subject* nas bases de dados MSRAction3D, MSRAction3DExt, UTKinect-Action, MSRDailyActivity3D e em uma base combinada composta por amostras das mesmas, onde desempenhos de 94,58%, 94,05%, 91,92%, 78,12% e 90,92% foram obtidas, respectivamente. O autor do trabalho indica que trabalhos futuros envolvem a adição da informação de esqueletos 3D e segmentação de objetos para aumentar o desempenho do reconhecedor.

Um reconhecedor que consolida os *frames* de um vídeo em uma estrutura única, denominada *Binary Motion Image* (BMI), para o reconhecimento de ações é proposto. Dado um vídeo de *input*, o fundo de cada um de seus *frames* é removido e eles são consolidados em uma única imagem, onde um peso é atribuído para cada *frame*. Por último, uma RNC processa a BMI para efetuar o processo de reconhecimento de ação. O reconhecedor também pode ser aplicado para mapas de profundidade e foi validado pelo métodos do *leve-one-out* na base de dados Weizmann, onde obteve 100% de desempenho, e pelos métodos *cross-subject* e validação cruzada na base de dados MSRAction3D, onde atingiu um desempenho de 98,5%. Trabalhos futuros envolvem a aplicação deste método em sistemas de vigilância onde o comportamento de múltiplas pessoas serão extraídos ao mesmo tempo.

Em Konda et al. (2015), um reconhecedor de ações online que utiliza descritores de movimentação aplicados a uma plataforma robótica é proposto. Uma sequência de *frames* é processada em um *autoencoder* denominado *synchrony autoencoder* (SAE), que descreve suas informações de movimentações e perspectiva. Uma base de dados foi criada pelo autor do trabalho, sendo ela composta por 574 amostras capturadas por 8 voluntários que realizaram 8 tipos de ações. O método de validação utilizado foi o *cross-subject* e um desempenho de 85,39% foi obtido. Além disso, o algoritmo também foi treinado em um ambiente online e um desempenho de 74,91% foi obtido.

Um método de reconhecimento de ações utilizando uma *Hierarchical Recursive Neural Network* (HRNN) é proposta em Yong Du, Wang e Wang (2015). Informações das juntas de um esqueleto 3D são separadas em cinco partes do corpo humano: braço direito, braço esquerdo, torso, perna direita e perna esquerda. Essas informações são consumidas por um conjunto de RNRs, que são responsáveis por criar descritores. Por último, a classificação é realizada

por uma Rede Neural Perceptron de uma camada. O método foi avaliado nas bases de dados MSRAction3D, MHAD e HDM05, onde mostrou alta eficiência computacional, além de obter desempenhos de 94,49%, 100% e 96,92%, respectivamente. Os métodos de validação utilizados foram o *cross-subject* para as bases MSRAction3D e MHAD e a validação cruzada para a base HDM05.

O reconhecimento de ações é aplicado em uma plataforma robótica de assistência doméstica em Papakostas et al. (2016). O método utilizado é composto uma RNC que descreve as hipóteses para as atividades realizadas em cada *frame* de um vídeo. Uma RNC pré-treinada na base de dados ILSVRC-2012 foi utilizada e uma base de dados gerada pelo autor do trabalho composta por 272 amostras coletadas por 12 voluntários e com 3 ações foi utilizada para sua validação. Os métodos utilizados foram a validação cruzada e a *cross-subject*, onde um desempenho de 95% foi obtido para ambos. Por último, é indicado que trabalhos futuros envolvem a utilização de informações de profundidade e áudio para melhorar o desempenho do reconhecimento de ações.

Em Jing Zhang et al. (2016), uma *survey* que busca avaliar as bases de dados utilizadas para o reconhecimento de ações e avaliar os reconhecedores propostos na literatura é realizada. As bases de dados estudadas foram divididas em três categorias: perspectiva única, múltiplas perspectivas e interações com dois ou mais indivíduos. Além disso, o estudo destacou informações tais como o ano de criação, modalidades de dados, número de classes, quantidade de voluntários, número de amostras e método de validação. Os resultados mostram que essas bases de dados são, em sua maior parte, simples e não representam a realidade corretamente. Também é possível verificar que a podem existir inconsistências ao comparar reconhecedores devido a falta de definições de protocolos claros para suas validações nas bases de dados da literatura e, apesar disso, os métodos de validação mais comuns são o *leave-one-out* e *cross-subject*.

Um método de reconhecimento de ações em tempo real que utiliza esqueletos 3D como *input* é proposto em Meshry, Hussein e Torki (2016). Os esqueletos 3D obtidos de um vídeo são transformados em *gesturelets*, que são descritores que representam suas informações de juntas e cinemática como direção e velocidade de movimento. Uma SVM com *kernel* linear é então aplicada em diversos intervalos do vídeo, onde uma ação pode ser detectada caso o intervalo com maior valor ultrapasse o limiar definido. Experimentos foram realizados nas bases de dados MSRAction3D e MSRC-12 por meio dos métodos *cross-subject* e validação cruzada, onde os f-scores obtidos foram 0,871 e 0,930.

Em P. Wang et al. (2016), um reconhecedor capaz de ser treinado em bases de dados pouco volumosas que utiliza 3 RNCs para processar mapas de profundidade é proposto. Dada uma sequência de *frames* de profundidade, um descritor, denominado *Weighted Hierarchical Depth Motion Maps* (WHDMM), que é uma representação 2D de seu *input*, é obtido. Assim, cada canal do WHDMM é inserido a um conjunto de 3RNCs responsáveis por gerar a classificação da amostra. O método foi avaliado nas bases de dados MSRAction3D, MSRAction3DExt, UTKinect-Action, MSRDailyActivity3D e em uma nova base composta pela junção das anteriores. O método de validação utilizado foi o *cross-subject* e os desempenhos obtidos para cada base de dados foram 100%, 100%, 90,91%, 85% e 91,56%, respectivamente.

Um método de reconhecimento de ações baseado em uma RNC com 2 fluxos de processamento é apresentado em Feichtenhofer, Pinz e Zisserman (2016). Os *frames* RGB de *input* são selecionados e pré-processados em dois tipos de estruturas: *frames* RGB e fluxos ópticos. Os *frames* RGB são alimentados à uma RNC que faz a análise espacial e os fluxos ópticos dos *frames* adjacentes são alimentados à uma outra RNC que faz a análise temporal. Por último, a saída de ambas as redes neurais sofrem fusão por uma camada convolucional 3D e são classificadas por uma rede *fully-connected*. Experimentos foram realizados nas bases de dados UCF101 e HMDB51 pelo método *cross-subject* e desempenhos de 93,5% e 69,2% foram obtidos.

Em Lee et al. (2017), um método de reconhecimento de ações baseado em LSTM, denominado *Temporal Sliding* LSTM (TL-LSTM), é proposto. Dada uma sequência de esqueletos 3D, suas juntas sofrem transformações geométricas para reduzir sua sensitividade relacionada a posição, orientação e escala. As diferenças entre os *frames* obtidos são calculadas e utilizadas como entrada para a TL-LSTM, que captura relações temporais curtas, médias e longas e, por último, classifica a entrada. Experimentos foram realizados nas bases de dados MSRAction3D, UTKinect-Action e NTU RGB+D pelo método *cross-subject*, onde desempenhos de 97,22%, 96,97% e 74,60% foram obtidos. Além disso, o método *cross-view* foi aplicado nas bases NTU RGB+D, Northwestern-UCLA e UWA3DII e os resultados mostram desempenhos de 82,56%, 89,22% e 75,6%, respectivamente.

A utilização de 3D *Residual Networks* (ResNets) para o reconhecimento de ações é avaliada em Hara, Kataoka e Satoh (2017). A utilização de ResNest é proposta como substituição para 3D CNNs, onde, devido ao alto tempo de treinamento de ResNets profundas, a arquitetura ResNet-34 foi utilizada. Assim, para cada 16 *frames* de um vídeo RGB, a ResNet-34 é aplicada

para a classificação da ação sendo realizada. Um experimento foi realizado na base de dado Kinetics pelo método da validação cruzada e o resultado indica um desempenho de 68,9%.

Um método representação de informações para o reconhecimento de ações, denominado POTION, baseado na junção de informações de movimentação e aparência dos *frames* de um vídeo é apresentado em Choutas et al. (2018). Um algoritmo de estimativa de pose é executado onde esqueletos 3D são obtidos e mapas de calor são gerados para cada *frame*. Por último, esses mapas de calor são combinados e o descritor POTION do vídeo é obtido. Assim, esse descritor pode ser alimentado a classificadores, como uma RNC, para a classificação do *input*. Experimentos foram realizados nas bases de dados JHMDB, HMDB51, UCF101 e Kinectis pela validação cruzada, onde os desempenhos de 85,5%, 80,9% e 98,2% foram obtidos, respectivamente. Trabalhos futuros envolvem a utilização de técnicas de *sliding window* para a aplicação do método em cenários com vídeos contínuos.

Em Pichao Wang et al. (2018a), um método de representação de trajetória de esqueletos 3D, denominado *Joint Trajectory Maps* (JTM), que codifica as informações espaciais e temporais de suas juntas em imagens 2D é proposto. Uma sequência de juntas 3D é analisada, onde as mesmas são projetadas em 3 planos ortogonais que serão utilizados para gerar 3 JTMs. Assim, 3 RNCs, uma para cada JTM, inicializadas com os mesmos parâmetros da AlexNet, são treinadas para a classificação da ação sendo realizada e suas saídas são combinadas para gerar a classificação final. Experimentos foram realizados nas bases de dados NTU RGB+D, MSRC-12 Kinect Gesture, G3D e UTD-MHAD pelos métodos do *cross-subject* e desempenhos de 76,27%, 94,86%, 96,02% e 87,90% foram obtidos, respectivamente. Além disso, a validação *cross-view* foi aplicada na base de dados NTU RGB+D obtendo 80,80% de desempenho. Trabalhos futuros envolvem a aplicação do método em ambientes de reconhecimento online.

Um modelo de representação de sequências de grafos dinâmicos de esqueletos 3D é proposto em Chaolong Li et al. (2018). Esse modelo codifica informações espaciais e temporais desses grafos dinâmicos por meio da aplicação de filtros convolucionais. Por último, as saídas dessas convoluções são estendidas em uma arquitetura profunda de classificação. Experimentos foram realizados nas bases de dados Florence 3D pelo método *leave-one-out*, HDM05 pela validação cruzada, *Large Scale Combined* (LSC) e NTU RGB+D pelo*cross-subject*, onde desempenhos de 99,07%, 85,29%, 88,1% e 86,28% foram obtidos. Além disso, desempenhos de 85,4% e 74,85% foram obtidos nas últimas duas bases de dados pelo método *cross-view*.

Em Jun Liu et al. (2018), uma adaptação de LSTM, denominada *Global Context-Aware Attention LSTM* (GCA-LSTM), que utiliza mecanismos de atenção para melhorar seu desem-

penho de reconhecimento de ações é proposta. O mecanismo de atenção utilizado tem como objetivo auxiliar o classificador a focar nas juntas consideradas mais informativas de um esqueleto 3D em cada *frame*. Além disso, um *Recurrent Attention Mechanism* foi criado buscando o melhoramento contínuo da seleção dessas juntas. O método foi avaliado pelo *cross-subject* nas bases de dados NTU RGB+D, SYSU-3D e MHAD, obtendo desempenhos de 76,1%, 78,6% e 100%, além de um desempenho de 84% pelo *cross-view* na base NTU RGB+D. O método *leave-one-out* também foi aplicado na base UT-Kinect com um resultado de 99% e, por último, um desempenho de 94,9% foi obtido na base SBU-Kinect Interaction pela validação cruzada.

Um reconhecedor, denominado *Deep Reinforcement Learning* (DPRL), que busca selecionar os esqueletos 3D mais informativos de uma sequência para realizar o reconhecimento de ações é proposto. Uma técnica de *deep reinforcement learning* é utilizada para o processo de seleção de *frames*, que ocorre de forma progressiva e leva em consideração a qualidade dos esqueletos 3D nesses *frames* e suas relações. Por último, uma GCN é aplicada para realizar os processos de treinamento e classificação. Experimentos foram realizados nas bases de dados NTU RGB+D, SYSU-3D e UT-Kinect e o reconhecdor foi capaz de atingir desempenhos de 89,8%, 76,9% e 98,5%, respectivamente. Os métodos de validação utilizados foram o *cross-subject*, *cross-view*, validação cruzada e *leave-one out*.

Em Yan, Xiong e Lin (2018), um método de reconhecimento de ações baseado em esqueletos dinâmicos, denominado *Spatial-Temporal Graph Convolutional Networks* (ST-GCN) é proposto. Uma representação genérica para uma sequência de esqueletos é obtida por meio de convoluções espaciais e temporais realizadas por uma GCN. Por último, essa representação é alimentada à um classificador SoftMax para determinar a ação sendo realizada. O reconhecedor foi avaliado na base de dados NTU RGB+D e obteve 81,5% e 88,3% de desempenho para os métodos *cross-subject* e *cross-view*, respectivamente. Além disso, a validação cruzada foi aplicada na base Kinetics, onde os resultados obtidos foram 30,7% e 52,8% para as precisões top-1 e top-5.

Um método de codificação de sequências de esqueletos 3D, denominado *Skeleton Optical Spectra* (SOS), utilizado para o reconhecimento de ações é proposto em Hou et al. (2018). Uma sequência de esqueletos 3D é codificada para uma representação de uma imagem no espaço de cores HSB por meio da representação de cada junta 3D em 3 planos ortogonais, gerando 3 imagens SOS. Por último, cada SOS é alimentada à uma RNC e a classificação final é obtida por meio da fusão de seus resultados. Experimentos foram realizados pelo método *cross-subject*

nas bases de dados MSRC12, Gaming 3D (G3D) e UTD-MHAD e desempenhos de 94,27%, 95,45% e 86,97% foram obtidos, respectivamente.

Em Si et al. (2018), um método de reconhecimento de ações por esqueletos 3D denominado *Spatial Reasoning and Temporal Stack Learnin*g (SR-TSL) é proposto. O método é composto por uma *Spatial Reasoning Network* (SRN) e uma *Temporal Stack Learning Network* (TSLN), que são responsáveis por obter informações estruturais espaciais de alto nível e dinâmicas em cada *frame*, respectivamente. Por último, um classificador softMax é utilizado para a classificação da ação sendo realizada. O reconhecedor foi avaliado pelos métodos da validação cruzada e *cross-subject* na base de dados SYSU 3D, onde obteve 80,7% e 81,9% de desempenho. Por último, os métodos *cross-subject* e *cross-view* foram aplicados na base de dados NTU RGB+D e desempenhos de de 84,8% e 92,4% foram obtidos.

Um reconhecedor de ações que utiliza uma RNC e uma *deep bidirectional* LSTM (DB-LSTM) em vídeos RGB é proposto em Ullah et al. (2018). Características do vídeo são extraídas por meio de uma RNC e são alimentadas à DB-LSTM, onde o processo de aprendizado ou classificação das relações na sequência de descritores é realizado. Experimentos foram realizados nas bases de dados UCF-101, Youtube 11 Actions e HMDB51 pelo método da validação cruzada, onde o método obteve 91,21%,, 92,84% e 87,64% de desempenho, respectivamente, em cenários em tempo real.

Em Tanfous, Drira e Amor (2018), um método de descrição de características para a trajetória de uma sequência de esqueletos 3D que pode ser utilizada para o reconhecimento de ações é proposto. Dada uma sequência de esqueletos 3D, sua trajetória é mapeada em um *Kendall's Shape Space* por meio de um dicionário de formas geométricas. Experimentos foram realizados para validar os descritores onde 2 classificadores foram utilizados: LSTM e SVM. As bases de dados Florence 3D, UTKinect-Action e MSR-Action3D foram utilizadas onde o método de avaliação escolhido foi o *cross-subject* e os resultados obtidos foram precisões de 93,04%, 96,89% e 91,84%, respectivamente.

Um algoritmo de avaliação de bases para o reconhecimento de ações, denominado RESOUND, é proposto em Li, Li e Vasconcelos (2018). O RESOUND avalia bases de dados de reconhecimento de ações por meio de métodos para mensurar a calibragem, viés de representação e viés de classes. Assim, uma nova base de dados, denominada Diving48, foi criada com o auxílio do RESOUND para diminuir seu viés. Essa base foi compara com outras 7 bases amplamente utilizadas na literatura, onde os resultados obtidos mostram uma menor quantidade de viés.

Em Chao Li et al. (2018), Um método de aprendizado de descritores para sequências de esqueletos 3D baseado em RNC é proposto. Cada junta do esqueleto 3D é tratada com um canal de entrada, onde uma RNC é utilizada para aprender as relações, espaciais e temporais, entre as mesmas, resultando em descritores globais para a sequência de esqueletos. Por último, a entrada é inserida em um modelo de classificação de 2 fluxos que é utilizado para a detecção e classificação da ação sendo realizada. Experimentos foram realizados nas bases de dados NTU RGB+D, SBU Kinect Interaction e PKU-MMD, onde os métodos *cross-view* e *cross-subject* foram utilizados para a validação e desempenhos de até 91,1%, 98,6% e 94,2% foram obtidos, respectivamente.

Um método que realiza estimativa de poses e reconhecimento de ações em vídeos RGB e que possibilita um treinamento *end-to-end* é proposto em Luvizon, Picard e Tabia (2018). Inicialmente, os *frames* RGB são consumidos por uma RNC para a estimativa da pose, onde um mapa de calor das possíveis posições das juntas é obtido. Por último, esses mapas são consumidos por duas RNCs, uma baseada em pose e outra baseada em características espaciais, onde suas saídas sofrem fusão e são alimentadas a uma camada *softmax* para a classificação final. Experimentos foram realizados nas bases de dados MPII Human Pose Dataset, Human3.6M, Penn Action e NTU RGB+D pelos métodos da validação cruzada e *cross-subject*, onde um desempenhos de até 98,6% e 85,5% foram obtidos.

Em Song et al. (2018), uma RNR de atenção baseada em LSTM com o objetivo de realizar a detecção e reconhecimento de ações é proposta. O método utiliza uma sequência de esqueletos 3D, onde um mecanismo de atenção atribui níveis de importância para cada junta desses esqueletos, além de propor os segmentos de frames onde há possibilidade da ocorrência de uma ação. Por último, esses seguimentos são processados por uma RNR baseada em LSTM para o reconhecimento das ações sendo realizadas. O método de validação *cross-subject* foi aplicado nas bases de dados SBU Kinect Interaction, NTU RGB+D e PKU-MMD e o resultados obtidos foram 91,5%, 79,4% e 86,9%. Além disso, o método *cross-view* foi aplicado nas duas últimas, onde desempenhos de 87,6% e 92,6% foram obtidos.

Um método de reconhecimento de ações em sequências de esqueletos 3D baseado em *clips* é proposto em Ke et al. (2018). As juntas dos esqueletos 3D são transformadas em 3 *clips*, cada um representando um canal de suas coordenadas. Os *clips* podem gerados à partir do posicionamento relativo das juntas ou pela rotação dos esqueletos 3D e, por último, são consumidos por uma *Multitask Convolutional Neural Nertwork* (MTCNN) para o processo de treinamento e classificação. Experimentos foram realizados na base de dados NTU RGB+D

pelos métodos *cross-subject* e *cross-view* e os resultados obtidos foram precisões de 81,09% e 87,37%, respectivamente. O método de validação *cross-subject* também foi aplicado nas bases Nothwestern-UCLA Multiview Action 3D e MHAD, onde os desempenhos obtidos foram 86,82% e 100%. Além disso, a validação cruzada foi aplicada nas bases SBU Kinect Interaction e CMU, obtendo 94,17% e 88,80% de desempenho e, por último, o método *leave-one-out* obteve 99% de desempenho na base de dados UTKinect-Action3D.

Em Pichao Wang et al. (2018b), um reconhecedor de ações que utiliza três descritores, adquiridos por meio de uma sequência de *frames* de profunidade, denominados *Dynamic Depth Image* (DDI), *Dynamic Depth Normal Images* (DDNI) e *Dynamic Depth Motion Normal Imagens* (DDMNI) é proposto. Primeiramente, os *frames* de profundidade de *input* são segmentados de acordo com o tempo inicial e final de cada ação. Logo após, os três descritores propostos são extraídos no formato de uma imagem dinâmica, onde cada uma é classificada por uma RNC. Por ultimo, a fusão dos resultados das 2 RNCs é realizada e o resultado final é obtido. Experimentos foram realizados pelo método *cross-subject* nas bases de dados ChaLearn LAP IsoGD e ChaLearn LAP ConGD, onde um desempenho de 59,21% foi obtido para a primeira e um *Jaccard Index* de 0,4109 foi obtido para a segunda. Por último, os métodos *cross-subject* e *cross-view* foram aplicados na base de dados NTU RGB+D, onde resultados mostram um desempenho de 87,08% e 84,22% ,respectivamente.

Uma RNC 3D com o objetivo de reconhecer ações em vídeos RGB por meio de 2 fluxos de processamento é proposta em Xuanhan Wang et al. (2018). O método possibilita a utilização de entradas de tamanho variável, uma vez que seus descritores de tamanho fixo são obtidos por meio de uma uma RNC de *pyramid pooling*. Assim, descritores espaciais e de movimentação são obtidos, sofrem fusão e, por último, são alimentados à um classificador que aprende suas relações temporais tal como uma LSTM. Experimentos foram realizados nas bases de dados UCF-101 e HMDB51 pelo método *cross-subject* onde desempenhos de 92,6% e 70,5% foram obtidos. Por último, a validação cruzada foi aplicada na base de dados ACT, obtendo um desempenho de 81,9%.

Um método de RNCs com multi-fluxos para o reconhecimento de ações em vídeos RGB é proposto em Tu et al. (2018). O método realiza a extração de *background* das imagens, além de extrair a região do ator que apresenta maior movimentação. Assim, descritores de movimentação e aparência da região de interesse são obtidos e alimentados à RNCs para o reconhecimento da ação sendo realizada. Experimentos pelo método *cross-subject* foram realizados nas bases de dados UCF101, JHMDB e HMDB51, onde o método proposto obteve desempenhos de 94,5%,

71,17% e 69,8%, respectivamente. Além disso, a validação cruzada foi aplicada na base UCF Sports onde um desempenho de 97,53% foi obtido. Novos trabalhos envolverão a melhoria do tempo computacional do método, uma vez que ele não é computacionalmente eficiente.

Em Kun Liu et al. (2018), um novo método de reconhecimento de ações em tempo real, denominado *Temporal Convolutional 3D Network* (T-C3D) é proposto. Os *frames* RGB são alimentados à T-C3D onde as características espaciais e temporais são obtidas, criando um descritor do vídeo que é consumido por uma camada *fully-connected* e, por último, a ação é classificada por uma camada *softmax*. O método de validação *cross-subject* foi aplicado nas bases de dados HMDB51 e UCF101, onde resultados apontam desempenhos de 92,5% e 62,8%, respectivamente. Trabalho futuros envolvem a adaptação da aplicação para ambientes online, onde o método receberá os *frames* incrementalmente ao invés dos vídeos por completo para análise.

Um método reconhecimento de ações em vídeos RGB baseado em uma *Hidden two-stream* CNN é proposto em Zhu et al. (2017). O método busca um maior desempenho computacional por meio da simplificação do cálculo do fluxo óptico. Assim, o fluxo óptico é obtido através de uma CNN não supervisionada, denominada *MotionNet* e, por último, essa informação sofre fusão com as informações espaciais do modelo para a classificação final do *input*. Experimentos foram realizados pelo método *cross-subject* nas bases de dados UCF101 e HMDB51, onde desempenhos de 97,1% e 78,7% foram obtidos. Além disso, a validação cruzada foi aplicada nas bases THUMOS14 e ActivityNet obtendo resultados de 80,6% e 91,2%.

Em Dong Li et al. (2019), um método de detecção e reconhecimento de ações que utiliza um mecanismo de atenção, denominado AttCell, para aumentar a efetividade de seu reconhecimento é proposto. Um vídeo de *input* é separado em modalidades tais como *frames* RGB e fluxo óptico. Para cada modalidade, aplica-se uma RNC de 1 camada com o objetivo de obter descritores. Assim, o STAN é aplicado para estimar as probabilidades de ocorrência de ações no vídeo e segmentá-lo de acordo com as mesmas. Por último, os descritores segmentados são alimentados a uma LSTM que destaca os mais indicativos e os utiliza para classificar a ação realizada no vídeo. Experimentos foram realizados nas bases de dados UCF-101, Columbia Consumer Videos (CCV) e THUMOS14, onde obteve um mAP de 93,6%, 83,7% e 77,3%, respectivamente. Os métodos de validação utilizados foram o *cross-subject* e validação cruzada.

Uma comparação entre 10 métodos de reconhecimento de ações que utilizam informações obtidas através de sensores Kinect é realizada em Wang, Huynh e Koniusz (2019). A comparação tem como objetivo avaliar a diferença entre o desempenho de métodos baseados

em descritores manuais e método que utilizam descritores baseados em *deeplearning*, além de avaliá-los por meio dos métodos do *cross-view* e *cross-subject*. Experimentos foram realizados nas bases de dados MSR-Action3D, 3D Action Pairs, CAD-60, UWA3D Activity Dataset, UWA3D Multiview Activity II e NTU RGB+D e os resultados obtidos mostram que os métodos que utilizam descritores baseados em *deeplearning* têm um desempenho maior para bases de dados volumosas, porém os que utilizam descritores manuais se mostram mais precisos para bases de dados não volumosas. Além disso, é possível notar que os descritores manuais não têm uma boa capacidade de generalização entre bases de dados diferentes. Por último, o estudo mostra que descritores baseados em esqueletos 3D apresentam um desempenho melhor para as validações pelos métodos *cross-view* e *cross-subject*.

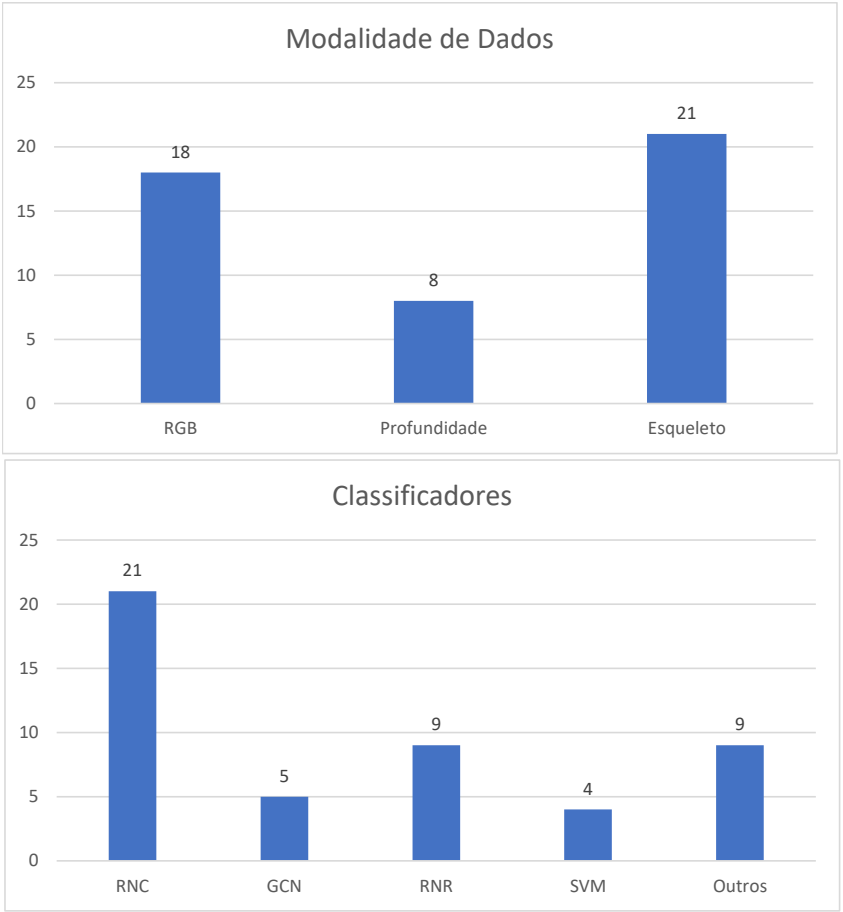
Em Hong-Bo Zhang et al. (2019), uma *survey* com o objetivo de apresentar o estado da arte dos métodos de reconhecimento de ações e detecção de ações, além de métodos de reconhecimento de interações humano-objeto é realizada. Para cada método estudado, informações tais como descritores, tipo de dados utilizados e representações de características são especificadas. O estudo mostra que os métodos baseados em *deeplearning* têm um melhor desempenho do que os métodos tradicionais, onde há duas principais metodologias utilizadas: métodos baseados em redes neurais convolucionais com 2 fluxos de processamento e métodos baseados em LSTM. Além disso, observa-se que estudos na analise de dados multimodais foi pouco explorada na área. Por último, o estudo destaca que os método de detecção de ações apresentam uma alta complexidade, o que dificulta sua utilização em aplicações online.

Um método de reconhecimento de ações utilizando esqueletos 3D é proposto em Ziyu Liu et al. (2020). A sequência de esqueletos 3D de entrada é consolidada em uma única estrutura de grafo e alimentada à uma GCN que realiza a classificação. Experimentos foram realizados nas bases NTU RGB+D 120 pelo método *cross-subject* e *cross-setup* e desempenhos de 86,9% e 88,4% foram obtidos. Além disso, outra validação foi realizada na base de dados Kinetics400 pelo método da validação cruzada, onde as precisões obtidas foram 38% para o top-1 e 60,9% para o top-5.

Por último, um método de reconhecimento de ações utilizando esqueletos 3D e *shift graph convolutions* é proposto em Cheng et al. (2020). O método é denominado Shift-GCN e realiza convoluções na estrutura de grafos do esqueleto 3D por meio de *shift convolutions* de forma que as convoluções em cada junta sejam não locais e considerem todas as outras juntas da estrutura. Experimentos foram realizados na base de dados NTU RGB+D pelos métodos *cross-subject* e *cross-view*, onde os resultados obtidos mostram precisões de 96,5% e 90,7%, respec-

tivamente. Além disso, a base NTU RGB+D 120 também foi utilizada para a validação onde desempenhos de 85,9% e 87,6% foram obtidos pelos métodos *cross-subject* e *cross-setup*. Por último, o *cross-view* foi aplicado na base Northwestern-UCLA, onde uma precisão de 94,6% foi obtida.

Figura 28 – Modalidades de dados e classificadores utilizados pelos reconhecedores de ações estudados.



Fonte: Autor.

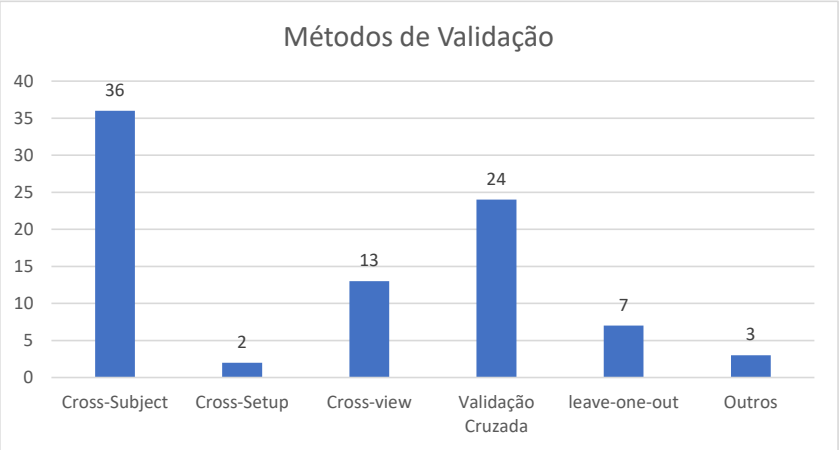
A Figura 28 resume as modalidades de dados e classificadores utilizados pelos trabalhos estudados. É possível observar que a modalidade mais utilizada é a de esqueletos 3D, sendo que grande parte desses reconhecedores são recentes tais como o MS-G3D (LIU, Z. et al., 2020) e o Shift-GCN (CHENG et al., 2020). Além disso, é possível observar que essa modalidade de dados é mais robusta em relação à perspectiva e *setup*, uma vez que os esqueletos 3D removem as informações tais como *background* e cor da roupa, porém informações espaciais como possíveis objetos utilizados nas imagens podem ser perdidas pelo mesmo motivo.

Também é possível observar que o classificador mais utilizado é a RNC, visto que ela foi utilizada pela maioria dos reconhecedores durante o período onde métodos de *deeplearning*

começaram a ser explorados e se mostrar mais eficientes. São bons exemplos desta metodologia métodos tais como Kay et al. (2017) e Pichao Wang et al. (2015). Por outro lado, métodos RNRs e GCNs mostraram um bom desempenho para obter relações temporais de seus *inputs*, logo é possível observar um aumento da utilização destes classificadores na literatura recente em trabalhos como Ziyu Liu et al. (2020) e Dong Li et al. (2019).

Assim, dado o objetivo do trabalho proposto, técnicas que utilizam esqueletos 3D como modalidades de dados serão escolhidas, uma vez que a extração desta estrutura pode ser realizada com baixo tempo computacional, além de demonstrar um melhor desempenho para cenários diferentes. Além disso, métodos que utilizam GCNs como classificadores tais como o MS-G3D (LIU, Z. et al., 2020) e o Shift-GCN (CHENG et al., 2020) serão utilizadas, em virtude de seu alto desempenho para o reconhecimento de bases volumosas e complexas como a NTU RGB+D 120.

Figura 29 – Métodos de validação utilizados pelos reconhecedores estudados.



Fonte: Autor.

Por último, a Figura 30 mostra os métodos de validação mais utilizados na literatura estudada, onde é possível observar que os métodos *cross-subject* e validação cruzada são os mais utilizados, sendo o primeiro utilizado em trabalhos que contêm as anotações dos usuários responsáveis por cada amostra e o segundo utilizado por trabalhos que utilizam amostras coletadas de plataformas da internet onde não há esse controle. Uma vez que este trabalho utilizará a base NTU RGB+D 120 (LIU, J. et al., 2019) para o treinamento dos reconhecedores, os métodos propostos pelos autores da base de dados serão utilizados, sendo eles o *cross-subject* e o *cross-setup*.

4 METODOLOGIA

Este Capítulo tem como objetivo apresentar a metodologia deste trabalho, que está dividida em duas etapas: obtenção das bases de dados e treinamento do reconhecedores. Assim, a Seção 4.1 busca descrever o método de escolha e obtenção das duas bases de dados utilizadas neste trabalho e a Seção 4.2 busca detalhar como os 12 reconhecedores foram obtidos e treinados a partir dessas bases de dados.

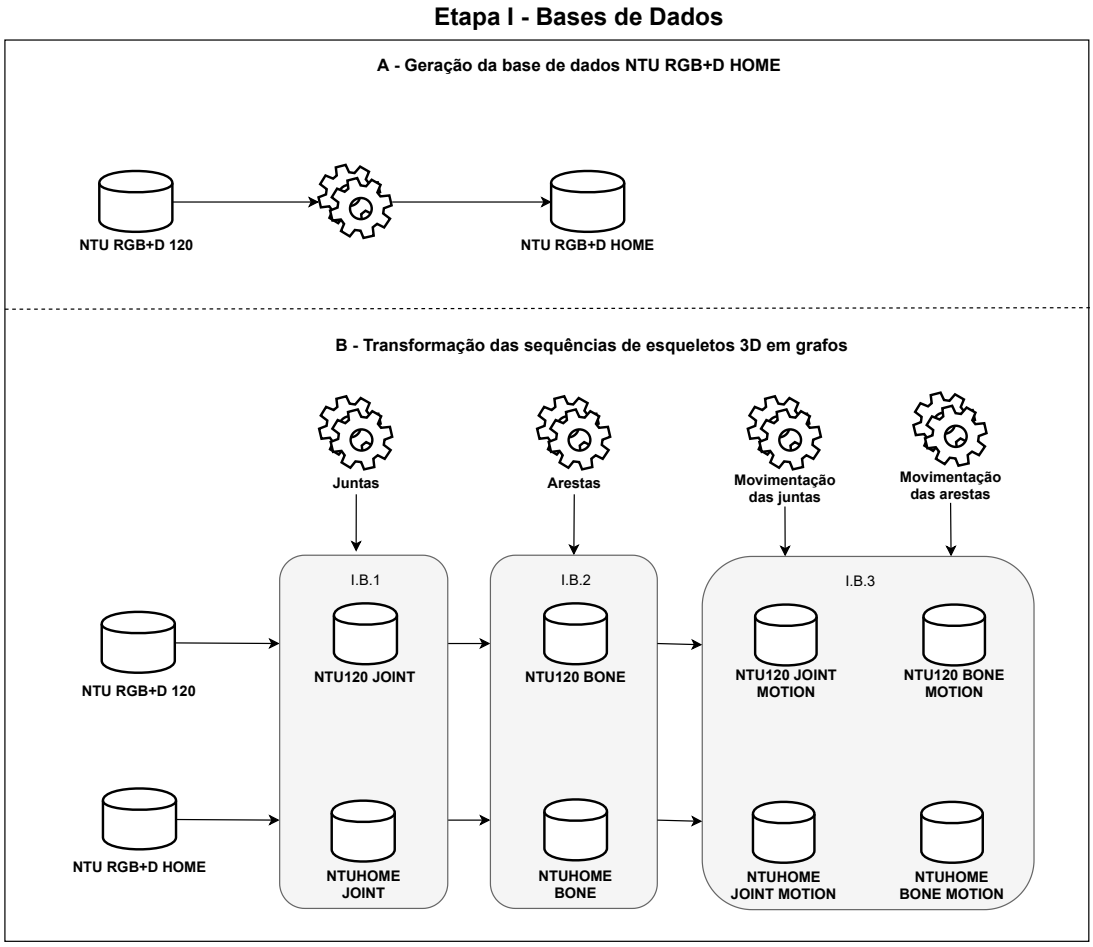
4.1 ETAPA I - BASES DE DADOS

A primeira etapa é a escolha de duas bases que representam ações sendo realizadas por um ser humano em seu dia-a-adia em um ambiente doméstico, uma vez que bases de dados volumosas e que representem o problema que deverá ser solucionado são necessárias para o treinamento dos reconhecedores. Assim, a Figura 30 descreve este processo, onde a base de dados NTU RGB120 foi escolhida, uma vez que mostra uma maior variedade de voluntários e perspectivas de coleta, além de disponibilizar uma quantidade volumosa de cerca de 114480 amostras. Por outro lado, para verificar o desempenho dos reconhecedores em um contexto menos complexo, um subconjunto de 23784 amostras, representando 25 classes, será extraído da base de dados NTU RGB+D 120, sendo esta nova base de dados denominada NTU RGB+D HOME. Por último, as amostras de ambas as bases devem ser processadas na estrutura de grafos espaciais e temporais utilizadas pelos reconhecedores MS-G3D e Shift-GCN. Assim, a Seção 4.1.1 descreve o processo de adaptação da NTU RGB+D 120 para a obtenção da base NTU RGB HOME e a Seção 4.1.2 descreve o processamento dessas bases em 4 modalidades de grafos para a alimentação dos reconhecedores.

4.1.1 Geração da base de dados NTU HOME (I.A)

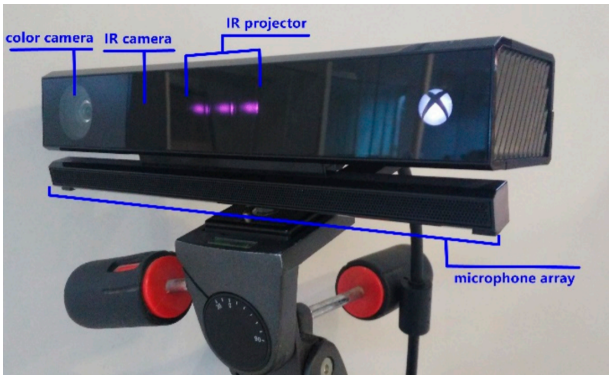
A base de dados NTU RGB+D 120 proposta em Jun Liu et al. (2019) é uma das maiores bases de vídeos de ações humanas RGB+D na literatura e pode ser encontrada em `http://rose1.ntu.edu.sg/datasets/actionrecognition.asp`. Ela foi coletada por 106 voluntários em 155 perspectivas diferentes, onde 114480 vídeos de 120 classes de ações se encontram disponíveis com 4 modalidades de dados: RGB, mapas de profundidade, esqueletos 3D e infravermelho.

Figura 30 – Metodologia para a obtenção das bases NTU RGB+D120 e NTU HOME.



Fonte: Autor.

Figura 31 – Sensor Microsoft Kinect V2.

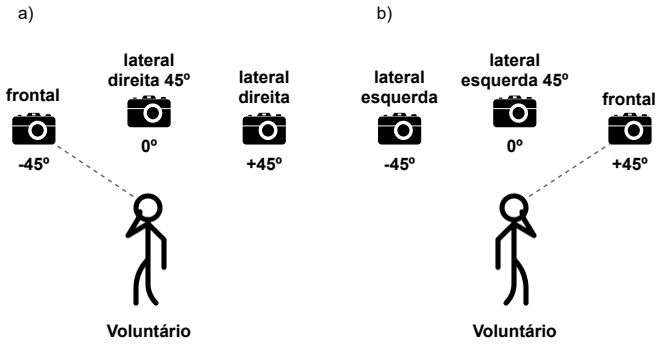


Fonte: (JIAO et al., 2017).

Os vídeos foram coletados por meio de um sensor Microsoft Kinect V2 (Figura 31) a 30 *frames* por segundo, onde as 155 perspectivas foram obtidas por meio de 31 *setups* diferentes com 5 perspectivas de coleta cada. A Tabela 2 descreve a altura e distância de posicionamento

das câmeras de coleta. Para cada *setup* apresentado, foi possível coletar 5 perspectivas diferentes por meio de 3 câmeras posicionadas em -45º, 0º e +45º dos voluntários. Assim, duas gravações foram realizadas por voluntário, onde a primeira gravação é executada com o voluntário voltado para a câmera da esquerda e outra voltado para a câmera da direita, totalizando as 5 perspectivas ilustradas pela Figura 32.

Figura 32 – Perspectivas de coleta para cada *setup* na base de dados NTU RGB+D 120. Existem 5 perspectivas para cada *setup*, sendo elas: frontal, lateral direita 45º, lateral direita, lateral esquerda 45º e lateral esquerda. O voluntário está voltado para a câmera da esquerda em (a), enquanto ele se encontra voltado para a câmera da direita em (b).



Fonte: Autor.

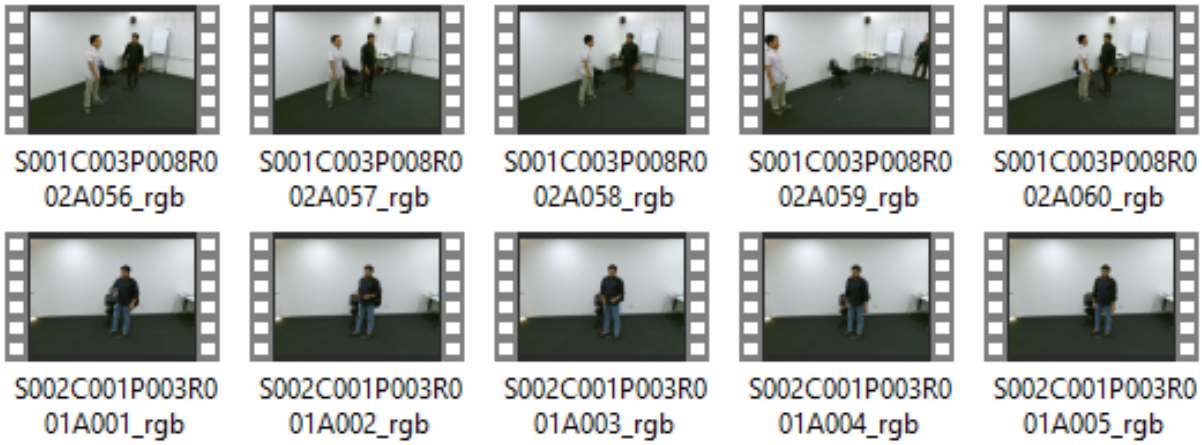
Tabela 2 – Descrição dos *setups* utilizados para a coleta dos vídeos da base NTU RGB+D 120. Vale notar que os *setups* 04 e 05 tem a mesma distância e altura da câmera, resultando nas mesmas perspectivas.

Setup	**Altura (m)**	**Distância (m)**	**Setup**	**Altura (m)**	**Distância (m)**
01	1.7	3.5	02	1.7	2.5
03	1.4	2.5	04	1.2	3.0
05	1.2	3.0	06	0.8	3.5
07	0.5	4.5	08	1.4	3.5
09	0.8	2.0	10	1.8	3.0
11	1.9	3.0	12	2.0	3.0
13	2.1	3.0	14	2.2	3.0
15	2.3	3.5	16	2.7	3.5
17	2.5	3.0	18	1.8	3.3
19	1.6	3.5	20	1.4	4.0
21	1.7	3.2	22	1.9	3.4
23	2.0	3.2	24	2.4	3.3
25	2.5	3.3	26	1.5	2.7
27	1.3	3.5	28	1.1	2.9
29	2.5	2.8	30	2.4	2.7
31	1.6	3.0	32	2.3	3.0

Fonte: (LIU, J. et al., 2019).

Para realizar o *download* dessa base de dados, deve-se preencher o formulário de requisição de acesso e, logo após o formulário sua aprovação, é possível obter as amostras da base de dados, que se encontram particionadas em 1 arquivo por *setup*, totalizando 32 arquivos. Esse procedimento deve ser repetido para cada modalidade de dados, uma vez que as amostras são armazenadas de forma independente para cada uma delas.

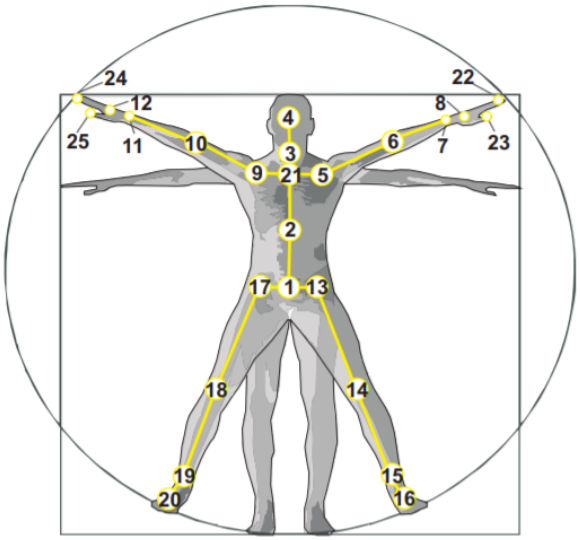
Figura 33 – Arquivos da base de dados NTU RGB+D 120.



Fonte: Autor

A Figura 33 mostra o exemplo da estrutura dos vídeos contidos na base de dados NTU RGB+D 120, onde as amostras de dados se encontram organizadas no padrão de nomenclatura **SSSSCCCCPPPPRRRRAAAA**, onde **SSSS** é o número identificador do *setup*, **CCCC** é o número da câmera, **PPPP** é o número do voluntário, **RRRR** é o número da gravação (duas por voluntário) e **AAAA** é a identificação da ação. Esta estrutura pode ser encontrada para todas as modalidades de dados, porém somente a modalidade de esqueletos 3D será utilizada, sendo suas amostras representadas pelas 25 juntas ilustradas pela Figura 34.

Figura 34 – As 25 juntas que são utilizadas para formar os esqueletos 3D na base de dados NTU RGB+D 120.



Fonte: (LIU, J. et al., 2019)

Apesar do grande volume da base NTU RGB+D 120, também é desejável verificar o desempenho dos reconhecedores em ambientes mais controlados, com foco doméstico, onde apenas uma parcela de suas ações seriam necessárias. Assim, um subconjunto de 25 classes dessa base foi escolhido para formar a nova base de dados denominada NTU RGB+D HOME. As classes escolhidas se encontram na Tabela 3 e tem como objetivo representar as principais ações que podem ocorrer em um ambiente doméstico ou ações relacionadas ao estado de saúde de um indivíduo.

Um *script* na linguagem Python foi criado para filtrar as amostras da NTU RGB+D 120 com base em sua nomenclatura e extrair somente as 25 classes desejadas. Por último, a base de dados NTU RGB+D HOME foi obtida, totalizando 23784 amostras, 25 classes, 106 voluntários e 155 perspectivas, assim como descrito pela Tabela 4.

Tabela 4 – Comparação entre as bases de dados NTU RGB+D 120 e NTU RGB+D HOME.

Base de Dados	**Vídeos**	**Classes**	**Voluntários**	**Perspectivas**
NTU RGB+D 120	114480	120	106	155
NTU RGB+D HOME	23784	25	106	155

Fonte: Autor.

Tabela 3 – Classes de ações selecionadas para a criação da base de dados NTU RGB+D HOME.

ID	ID NTU RGB+D 120	Ação
1	A1	Beber água
2	A2	Comer uma refeição
3	A3	Escovar os dentes
4	A8	Sentar
5	A9	Levantar
6	A11	Ler
7	A12	Escrever
8	A27	Pular
9	A28	Falar no telefone
10	A32	Tirar uma *selfie*
11	A37	Lavar o rosto
12	A41	Espirrar/tossir
13	A43	Cair no chão
14	A44	Sentir dor de cabeça
15	A45	Sentir dor no peito
16	A46	Sentir dor nas costas
17	A47	Sentir dor no pescoço
18	A48	Sentir náusea
19	A69	Polegar para cima
20	A70	Polegar para baixo
21	A74	Contar dinheiro
22	A85	Aplicar creme no rosto
23	A103	Bocejar
24	A104	Alongar
25	A105	Assoar o nariz

Fonte: Autor.

4.1.2 Transformação das sequências de esqueletos 3D em grafos (I.B)

Uma vez que os esqueletos 3D foram obtidos, deve-se transformá-los em estruturas que possam ser consumidas por uma GCN. Assim, os dados coletados para cada amostra podem ser utilizados para extrair informações sobre as movimentações voluntários, onde a Seção 4.1.2.1 descreve a utilização do posicionamento das juntas, a Seção 4.1.2.2 detalha a utilização das arestas que ligam essas juntas e a Seção 4.1.2.3 mostra como utilizar a movimentação de juntas ou arestas entre *frames* adjacentes.

4.1.2.1 Juntas 3D (I.B.1)

Uma das possíveis representações de um esqueleto 3D é transformá-lo em um grafo de acordo com os atributos de suas juntas. Nas bases de dados utilizadas neste trabalho, cada amostra foi armazenada em um arquivo com a extensão *skeleton*, onde se encontram especificados os esqueletos de até 2 indivíduos para cada *frame* do vídeo, além de conter as informações de posicionamento e orientação de cada junta.

Para cada esqueleto 3D, um tensor descritivo com o formato $C \times T \times V \times M$ é formado, sendo C a quantidade de canais (eixos de posicionamento das juntas), T a quantidade de *frames*, V a quantidade de juntas e M a quantidade máxima de esqueletos de voluntários. De acordo com as configurações do projeto proposto, o tensor apresentou um tamanho $3 \times 300 \times 25 \times 2$. Esse procedimento é realizado para as N amostras da base de dados e, por último, elas são armazenadas em um tensor com tamanho $N \times T \times V \times M$, que será utilizado pelo reconhecedor. Assim, esse procedimento é aplicado para ambas as bases de dados e 2 novas bases são obtidas: NTU120-JOINT e NTUHOME-JOINT.

4.1.2.2 Arestas (I.B.2)

Ao invés de utilizas as informações de cada junta do esqueleto 3D, também é possível utilizar as informações das arestas que as ligam. Para gerar esta representação, é necessária a geração prévia da representação de juntas descrita na Seção 4.1.2.1. Esses dados são carregados e então as arestas são obtidas pela subtração do posicionamento das juntas adjacentes. Por exemplo, a aresta $B_{1,2}$, que representa a ligação entre as juntas v_1 e v_2, pode ser obtida pela Equação 12. Essas informações são extraídas para ambas as bases de dados selecionadas e 2 novas bases de dados são obtidas: NTU120 BONE e NTUHOME BONE.

$$B_{1,2} = v_1 - v_2 \tag{12}$$

4.1.2.3 Movimentação (I.B.3)

Por último, outra forma de representar as informações do esqueleto 3D é a movimentação entre juntas ou suas ligações. Assim, dado um conjunto de informações de juntas (Seção 4.1.2.1) ou arestas (Seção 4.1.2.2), pode-se obter a informação suas movimentações entre *fra-*

mes ao invés de somente suas posições. A Equação 13 mostra um exemplo para o o vértice v onde a movimentação B_t é calculada para os *frames* adjacentes nos tempos t e $t + 1$. Ao aplicar esse procedimento nas bases de dados selecionadas, as bases NTU120 JOINT MOTION, NTU120 BONE MOTION, NTUHOME JOINT MOTION e NTUHOME BONE MOTION são obtidas.

$$B_t = v_{t+1} - v_t \tag{13}$$

4.2 ETAPA II - RECONHECEDORES

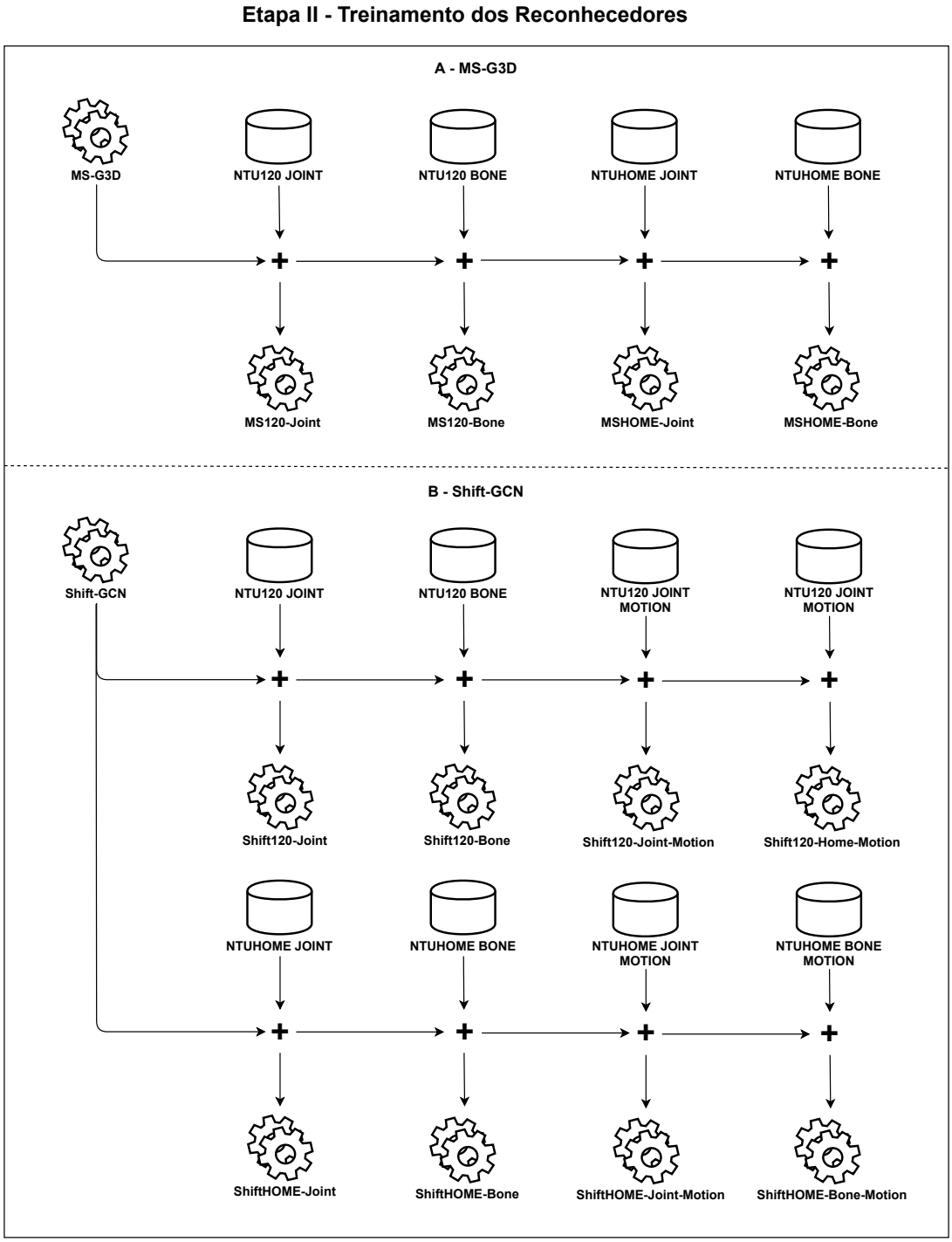
Após a escolha das bases de dados e processamento de suas amostras em estruturas de grafos, o treinamento dos reconhecedores deve ser realizado. Neste trabalho, serão treinados 6 modelos que serão descritos e seus parâmetros de treinamento detalhados nesta Seção. A Figura 35 ilustra os modelos propostos, onde a Seção 4.2.1 descreve os 2 modelos gerados a partir do reconhecedor MS-G3D e a Seção 4.2.2 descreve os 4 modelos gerados a partir do reconhecedor Shift-GCN.

4.2.1 MS-G3D (II.A)

O reconhecedor MS-G3D é um reconhecedor de ações que utiliza o modelo de uma GCN e esta Seção busca descrever seu processo de treinamento. O treinamento ocorrerá para as bases de dados NTU RGB+D 120 e NTU RGB+D HOME, que se encontram descritas na Seção 4.1. O MS-G3D pode consumir 2 tipos de dados, sendo eles as informações de juntas (Seção 4.1.2.1) e as informações de arestas (4.1.2.2), que resultaram nos seguintes modelos:

a) **MS120-Joint**: o reconhecedor MS-G3D treinado com as informações de juntas da base NTU120 JOINT;
b) **MS120-Bone**: o reconhecedor MS-G3D treinado com as informações de arestas da base NTU120 BONE;
c) **MSHOME-Joint**: o reconhecedor MS-G3D treinado com as informações de juntas da base NTUHOME JOINT;
d) **MSHOME-Bone**: o reconhecedor MS-G3D treinado com as informações de arestas da base NTUHOME BONE.

Figura 35 – Modelos de reconhecedores propostos.



Fonte: Autor.

O treinamento foi realizado em 2 placas de vídeos NVIDIA Tesla v100 e os reconhecedores foram treinados durante 60 épocas com um tamanho de *batch* de 32. Por último, os parâmetros seguiram as diretrizes em Ziyu Liu et al. (2020), sendo a taxa de aprendizagem es-

colhida $0{,}05$ que é divida por 10 nas épocas 30 e 50 e o otimizador utilizado foi o *Stochastic Gradient Descent* (SDG) com *momentum* de Nesterov.

4.2.2 Shift-GCN (II.B)

O reconhecedor Shift-GCN é um reconhecedor de ações que utiliza o modelo de uma GCN e esta Seção busca descrever seu processo de treinamento. O treinamento ocorrerá para as bases de dados NTU RGB+D 120 e NTU RGB+D HOME, que se encontram descritas na Seção 4.1. O Shift-GCN pode consumir 4 tipos de dados, sendo eles as informações de juntas (Seção 4.1.2.1), as informações de arestas (Seção 4.1.2.2) e a movimentação das juntas e das arestas (Seção 4.1.2.3), que resultaram nos seguintes modelos:

a) **Shift120-Joint**: o reconhecedor Shift-GCN treinado com as informações de juntas da base NTU120 JOINT;
b) **Shift120-Bone**:o reconhecedor Shift-GCN treinado com as informações de arestas da base NTU120 BONE;
c) **Shift120-Joint-Motion**: o reconhecedor Shift-GCN treinado com as movimentações de juntas da base NTU120 JOINT MOTION;
d) **Shift120-Bone-Motion**: o reconhecedor Shift-GCN treinado com as movimentações de arestas da base NTU120 BONE MOTION;
e) **ShiftHOME-Joint**: o reconhecedor Shift-GCN treinado com as informações de juntas da base NTUHOME JOINT;
f) **ShiftHOME-Bone**: o reconhecedor Shift-GCN treinado com as informações de arestas da base NTUHOME BONE;
g) **ShiftHOME-Joint-Motion**: o reconhecedor Shift-GCN treinado com as movimentações de juntas da base NTUHOME JOINT MOTION;
h) **ShiftHOME-Bone-Motion**: o reconhecedor Shift-GCN treinado com as movimentações de arestas da base NTUHOME BONE MOTION.

O treinamento foi realizado em 2 placas de vídeos NVIDIA Tesla v100 e os reconhecedores foram treinados durante 140 épocas com um tamanho de *batch* de 64. Por último, os paramêtros seguiram as direztrizes em Cheng et al. (2020), sendo a taxa de aprendizagem $0{,}1$ que é divida por 10 nas épocas 60, 80 e 100 e o otimizador utilizado foi o SDG com *momentum* de Nesterov.

5 EXPERIMENTOS

Este Capítulo tem como objetivo apresentar e discutir os experimentos que serão realizados para verificar e validar o desempenho dos 12 modelos de reconhecedores obtidos na metodologia deste trabalho (Capítulo 4). Para cada base de dados, os métodos de validação *cross-subject* e *cross-setup* serão aplicados, assim como proposto em Jun Liu et al. (2019), Além disso, vale notar que os particionamentos utilizados serão aqueles propostos pelos autores da bases para possibilitar a comparação entre os modelos, apesar de que a utilização de uma partição de testes é desejável em trabalhos futuros. Assim, a Seção 5.1 detalha os resultados dos reconhecedores treinados na base de dados NTU RGB+D 120 e a Seção 5.2 descreve os resultados dos reconhecedores treinados na base de dados NTU RGB+D HOME.

5.1 NTU RGB+D 120

Para verificar o desempenhos dos reconhecedores escolhidos e comparar seus desempenhos em relação aos artigos onde são propostos, a base de dados NTU RGB+D 120 foi escolhida. Os métodos de validação *cross-subject* e *cross-setup* foram selecionados e executados de acordo com as recomendações propostas em Jun Liu et al. (2019).

O primeiro método aplicado foi o *cross-subject*, onde as amostras são particionadas de acordo com o voluntário que as coletou. O particionamento adotado para o conjunto de treinamento é composto por 63561 amostras dos voluntários 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38, 45, 46, 47, 49, 50, 52, 53, 54, 55, 56, 57, 58, 59, 70, 74, 78, 80, 81, 82, 83, 84, 85, 86, 89, 91, 92, 93, 94, 95, 97, 98, 100, 103 e a partição de validação foi consolidada pelas 50919 amostras restantes. Além disso, vale ressaltar que o método é composto por um conjunto de treinamento e outro de validação e, apesar da aplicação de um terceiro conjunto de teste ser interessante para uma validação mais robusta, somente esses dois conjuntos são considerados em busca de manter a comparabilidade dos resultados dos reconhecedores neste trabalho com os resultados de seus trabalhos originais.

O segundo método aplicado foi o *cross-setup*, que busca verificar a robustez dos reconhecedores em relação a alterações do ambiente e perspectiva de coleta. Assim, o particionamento adotado para o conjunto de treinamento é composto por 55003 amostras dos *setups* de número par para treinamento e as 59477 amostras dos *setups* de número ímpar para validação, totalizando 16 *setups* para cada conjunto.

Por último, o desempenho dos reconhecedores foi obtido por meio da precisão dos mesmos de acordo com o particionamento dos métodos *cross-subject* e *cross-setup*. Assim, para cada reconhecedor, a precisão Top 1 representa a proporção de classificações corretas e a precisão Top 5 representa a proporção de amostras onde a classificação esperada se encontra entre as 5 maiores probabilidades.

5.2 NTU RGB+D HOME

Este experimento tem como objetivo avaliar os reconhecedores em um cenário com uma quantidade menor de classes em relação à base de dados NTU RGB+D 120 e com o foco nas principais atividades realizadas em um ambiente doméstico e relacionadas a saúde. Assim, uma vez que a base de dados NTU RGB+D HOME é composta por um subconjunto da NTU RGB+D 120, os reconhecedores treinados nela serão avaliados pelos métodos de validação *cross-suject* e *cross-setup* de acordo com as recomendações em Jun Liu et al. (2019).

A aplicação do método *cross-subject* foi realizada para verificar a robustez do reconhecedores em relação a diferentes voluntários. Assim, o conjunto de treinamento foi composto por 14818 amostras que foram coletadas pelos voluntários 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38, 45, 46, 47, 49, 50, 52, 53, 54, 55, 56, 57, 58, 59, 70, 74, 78, 80, 81, 82, 83, 84, 85, 86, 89, 91, 92, 93, 94, 95, 97, 98, 100 e 103. Por outro lado, o conjunto de validação é composto por 8966 amostras coletadas pelos voluntários restantes.

Em contraste, para verificar a robustez do reconhecedores em relação a diferentes ambientes e perspectivas de coleta, o método *cross-setup* foi aplicado. O particionamento adotado para o conjunto de treinamento é composto por 11325 amostras dos *setups* de número par e o conjunto de validação é composto por 12459 amostras dos *setups* de número ímpar, totalizando 16 *setups* para cada partição.

Vale notar que, assim como o protocolo proposto no Experimento da Seção 5.1, este experimento não utilizará o particionamento da amostras em conjunto de testes, uma vez que a comparação com os resultados originais dos reconhecedores é desejável. Além disso, para cada método de validação, a proporção de classificações corretas é obtida pela precisão Top 1 e a precisão Top 5 indica a proporção de amostras onde a classe esperada se encontra nas 5 maiores probabilidades de reconhecimento.

6 RESULTADOS

Este Capítulo tem com objetivo descrever os resultados obtidos através da aplicação dos experimentos propostos na Seção 5. Para cada base de dados, os métodos *cross-subject* e *cross-setup* foram aplicados e os resultados obtidos serão descritos e discutidos nas Seções a seguir, onde Seção 6.1 detalha os resultados do experimento realizado na base de dados NTU RGB+D 120, a Seção 6.2 descreve os resultados obtidos pela validação dos reconhecedores na base de dados NTU RGB+D HOME e a Seção 6.3 discute sobre os resultados obtidos em ambos os experimentos.

6.1 NTU RGB+D 120

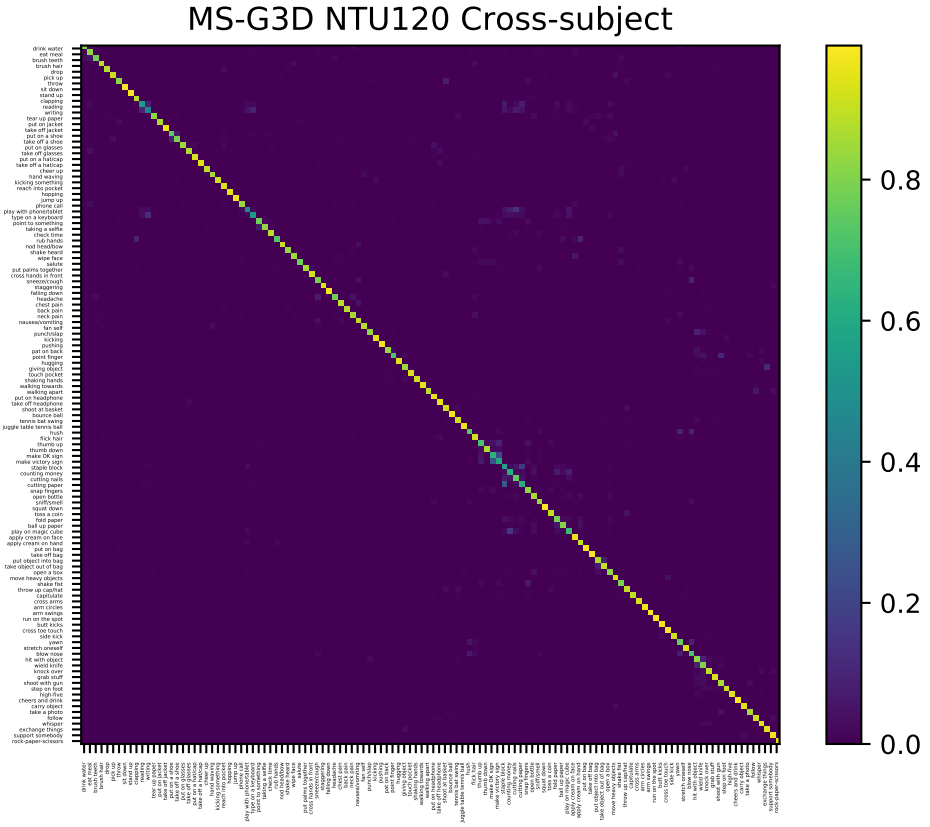
Esta Seção busca descrever sobre os resultados obtidos pelo experimento na base de dados NTU RGB+D 120 descrito anteriormente na Seção 5.1. A Tabela 5 ilustra os resultados para o primeiro experimento pelo método de validação *cross-subject* e é possível observar que, para cada tipo de informação utilizada, os desempenhos obtidos por todos os reconhecedores para a precisão Top 1 se encontram entre 79,03% e 84,63%, sendo esses resultados próximos aos originalmente reportados nas propostas dos métodos. Além disso, é possível observar que os métodos treinados com as informações de arestas mostraram um maior desempenho em relação aos outros formatos, sendo essa diferença próxima de 2,5% de precisão.

Tabela 5 – Resultados dos reconhecedores pelo método de validação *cross-subject* na base NTU RGB+D 120.

	MS-G3D		Shift-GCN	
Dados	**Top 1**	**Top 5**	**Top 1**	**Top 5**
Joint	81,96%	95,49%	81,63%	95,65%
Bone	84,63%	96,70%	83,42%	96,71%
Joint Motion	-	-	79,03%	94,62%
Bone Motion	-	-	80,30%	95,50%
Fusão	85,98%	97,09%	86,11%	97,52%

Fonte: Autor.

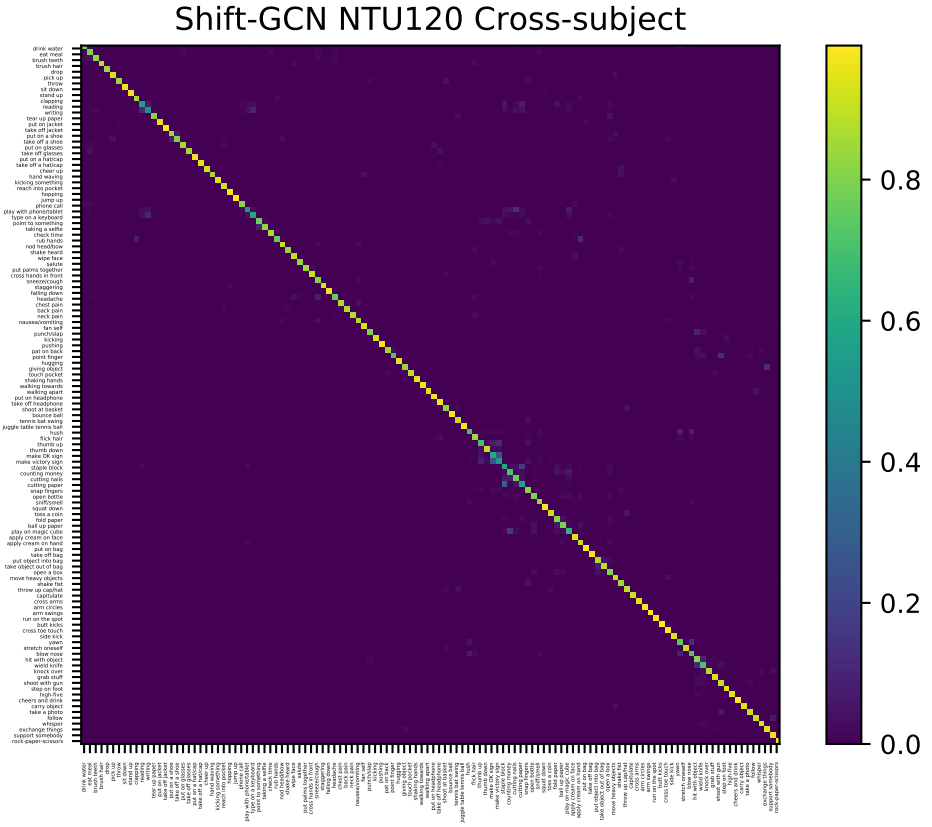
Figura 36 – Matriz de confusão obtida pelo MS-G3D na validação *cross-subject* na base NTU RGB+D 120.



Fonte: Autor.

Também é possível verificar um desempenho acima de 94,62% para a precisão Top 5, o que indica que, mesmo quando a classe com maior probabilidade da amostra não era a esperada, os reconhecedores apresentaram uma capacidade de atribuí-la entre as 5 maiores probabilidades. A fusão dos resultados do reconhecedores mostrou um aumento de desempenho do reconhecimento, onde houve uma melhora de 1,35% e 2,69% para os modelos derivados do MS-G3D e o Shift-GCN, respectivamente. Por último, as Figuras 36 e 37 ilustram as matrizes de confusão para os resultados desse procedinhmento e é possível observar que as predições ocorreram com desempenho estado da arte para a maioria das ações, exceto para ações semelhantes tais como ler e escrever ou dedão para cima e dedão para baixo.

Figura 37 – Matriz de confusão obtida pelo Shift-GCN na validação *cross-subject* na base NTU RGB+D 120.



Fonte: Autor.

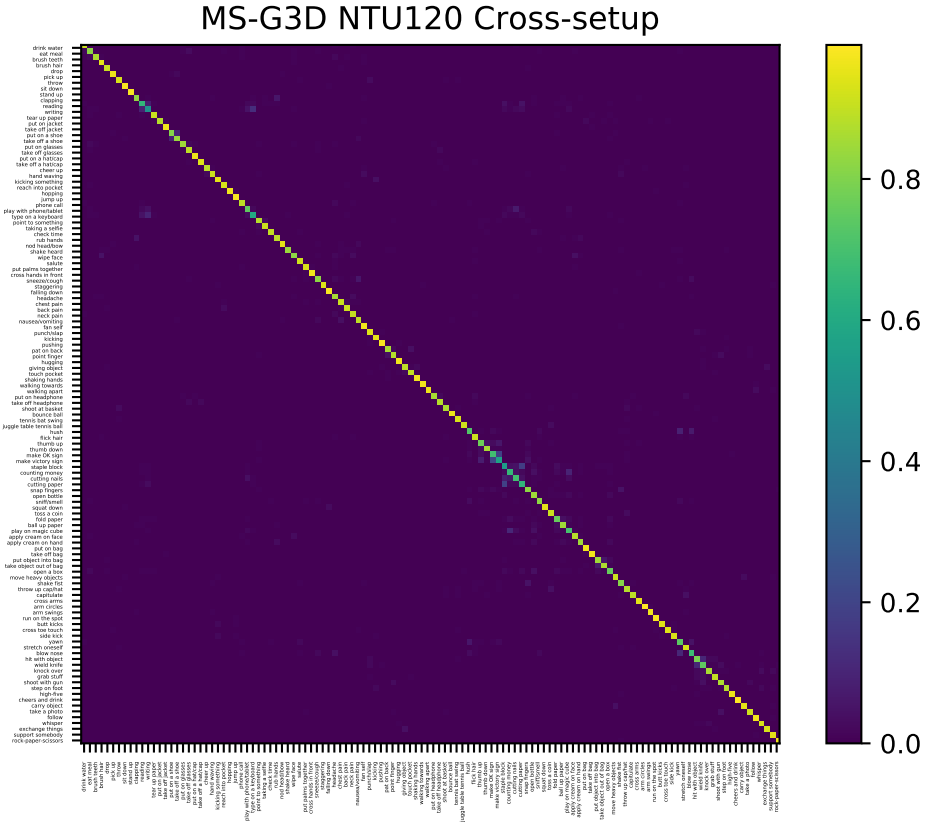
O segundo método aplicado foi o *cross-setup* e seus resultados se encontram na Tabela 6. É possível observar que as precisões Top 1 dos reconhecedores se encontram entre 80,95% e 85,80% e de acordo com os desempenhos reportados nas propostas originais dos métodos. Neste caso, as informações de arestas apresentaram um desempenho melhor para os reconhecedores derivados do MS-G3D e as informações de movimentação de juntas se mostraram mais eficientes para os reconhecedores derivados do Shift-GCN.

Tabela 6 – Resultados dos reconhecedores pelo método de validação *cross-setup* na base NTU RGB+D 120.

	MS-G3D		Shift-GCN	
Dados	**Top 1**	**Top 5**	**Top 1**	**Top 5**
Joint	83,68%	95,56%	83,06%	95,53%
Bone	85,80%	96,77%	80,95%	94,83%
Joint Motion	-	-	85,13%	96,57%
Bone Motion	-	-	81,82%	95,65%
Fusão	87,51%	97,23%	87,84%	97,39%

Fonte: Autor.

Figura 38 – Matriz de confusão obtida pelo MS-G3D na validação *cross-setup* na base NTU RGB+D 120.

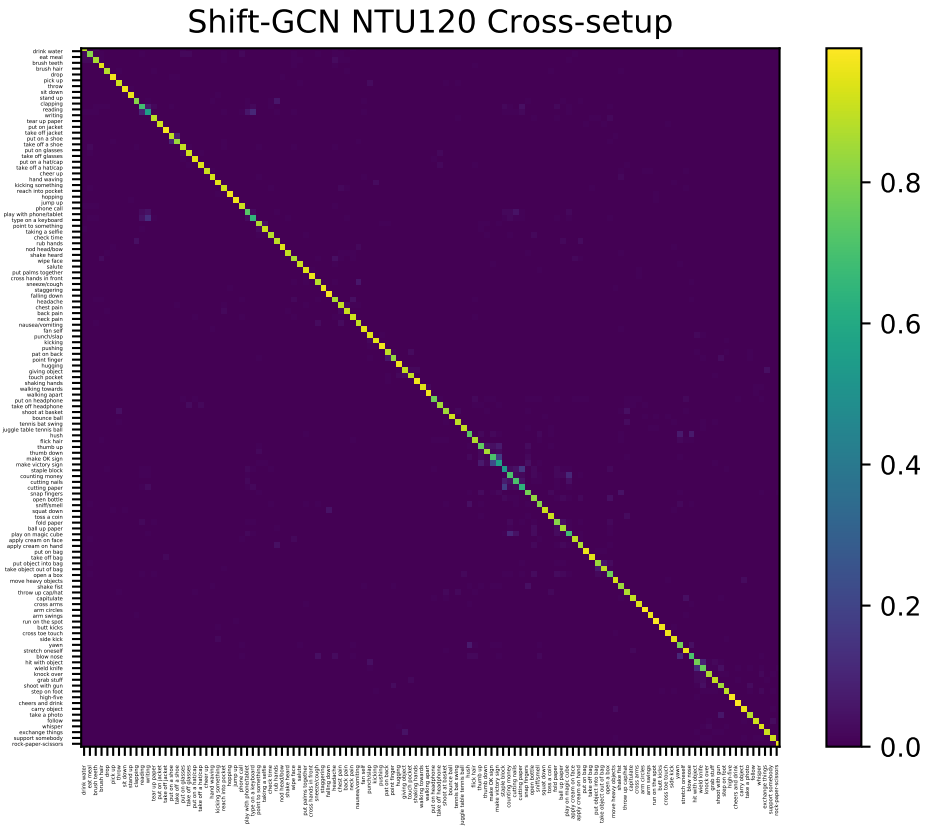


Fonte: Autor.

Além disso, os reconhecedores apresentaram uma robustez semelhante ao método do *cross-subject*, onde as precisões Top 5 se encontram entre 94,83% e 96,77%. Observa-se que a fusão dos resultados dos reconhecedores também resultou em um aumento de desempenho em relação a *setups* diferentes, sendo este de 1,71% e 2,71% para aqueles derivados do MS-

G3D e Shift-GCN, respectivamente. Por último, as matrizes de confusão desse procedimento se encontram ilustradas pelas Figuras 38 e 39 e foi possível observar que, da mesma maneira que no método do *cross-subject*, os menores desempenhos foram obtidos por classes compostas por movimentos semelhantes tais como ler e escrever ou dedão para cima e dedão para baixo. Neste caso, para os dois conjuntos de ações, o corpo do voluntário realiza as mesmas movimentações, sendo que as diferenças encontram-se em gestos realizados somente pelas mãos.

Figura 39 – Matriz de confusão obtida pelo Shift-GCN na validação *cross-setup* na base NTU RGB+D 120.



Fonte: Autor.

6.2 NTU RGB+D HOME

Esta Seção tem como objetivo detalhar os resultados obtidos pelos experimentos realizados na base de dados NTU RGB+D HOME (Seção 5.2). Assim, a Tabela 7 mostra os resultados obtidos para os reconhecedores treinandos nesta base de dados e avaliados pelo método *cross-subject*. Observa-se que os métodos apresentaram precisões Top 1 entre 80,11% e

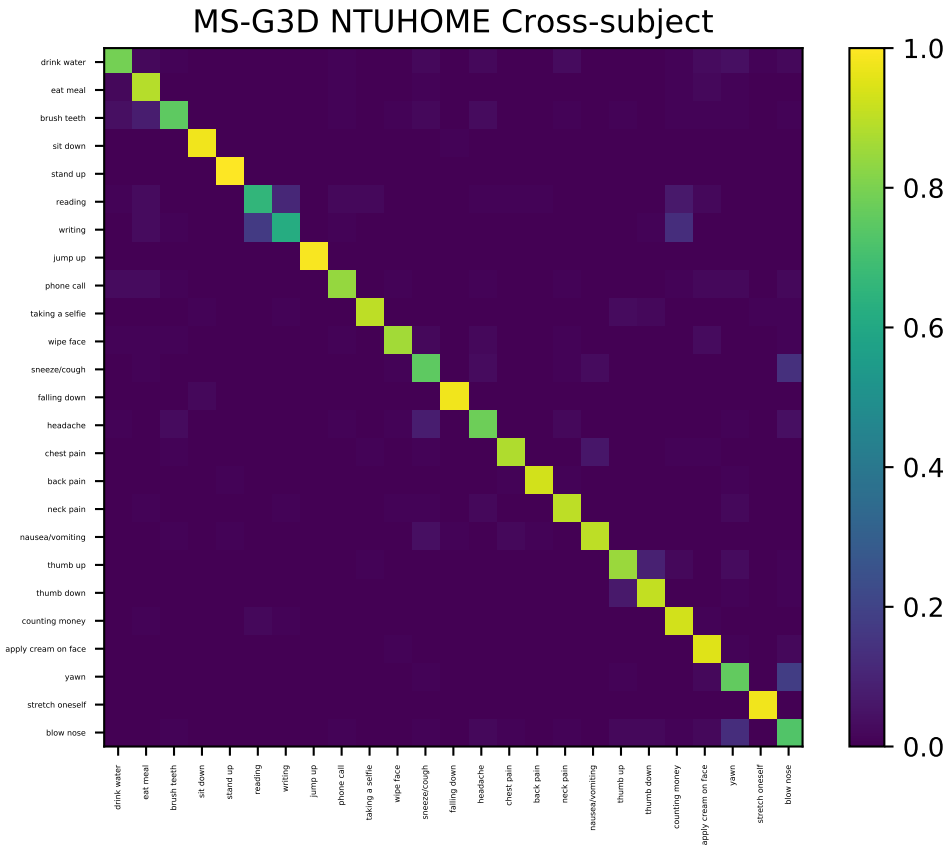
84,32% e que, neste caso, é possível verificar que as informações de arestas apresentaram um maior desempenho para os reconhecedores derivados do MS-G3D, assim como os resultados apresentados na Seção 6.1, e as informações de movimentação de juntas mostraram um maior desempenho para os reconhecedores do Shift-GCN.

Tabela 7 – Resultados dos reconhecedores pelo método de validação *cross-subject* na base NTU RGB+D HOME.

	MS-G3D		Shift-GCN	
Dados	**Top 1**	**Top 5**	**Top 1**	**Top 5**
Joint	83,46%	97,14%	80,11%	96,52%
Bone	84,32%	97,22%	80,36%	96,92%
Joint Motion	-	-	81,17%	97,02%
Bone Motion	-	-	80,76%	97,03%
Fusão	85,99%	97,75%	85,36%	98,09%

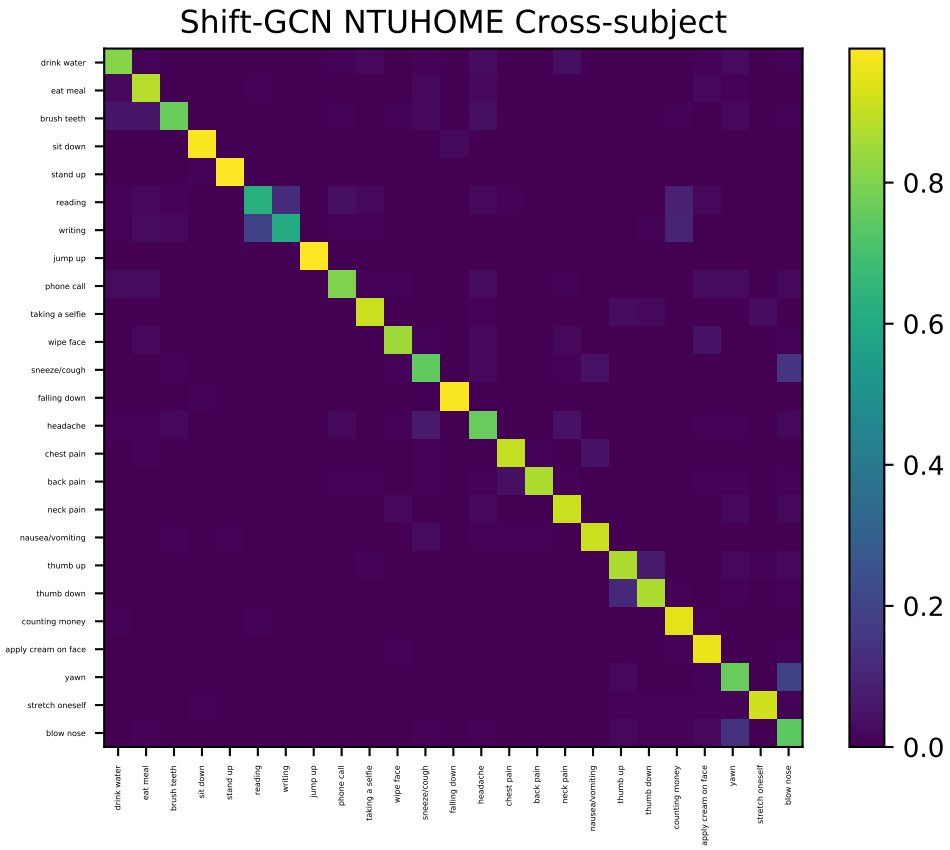
Fonte: Autor.

Figura 40 – Matriz de confusão obtida pelo MS-G3D na validação *cross-subject* na base NTU RGB+D HOME.



Fonte: Autor.

Figura 41 – Matriz de confusão obtida pelo Shift-GCN na validação *cross-subject* na base NTU RGB+D HOME.



Fonte: Autor.

Também é possível observar desempenhos acima de 96,52% das precisões Top 5 e que todos os reconhecedores mostraram robustez durante o processo de classificação para amostras de voluntários diferentes. A fusão dos resultados dos modelos mostrou uma melhora no processo de reconhecimento, onde um aumento de 1,67% e 4,19% foi observado para os reconhecedores derivados do MS-G3D e Shift-GCN, respectivamente. Por último, as Figuras 40 e 41 ilustram as matrizes de confusão do procedimento de fusão e foi possível observar características semelhantes ao Experimento 1, onde as classes com movimentações semelhantes tais como ler e escrever apresentaram um desempenho reduzido e de difícil diferenciação pelos reconhecedores.

A Tabela 8 mostra os resultados para a avaliação pelo método *cross-setup*. As precisões Top 1 obtidas pelos reconhecedores apresentam robustez em relação a ambientes e perspetivas diferentes, sendo seus melhores desempenhos de 83,75% e 81,47% para os reconhecedores

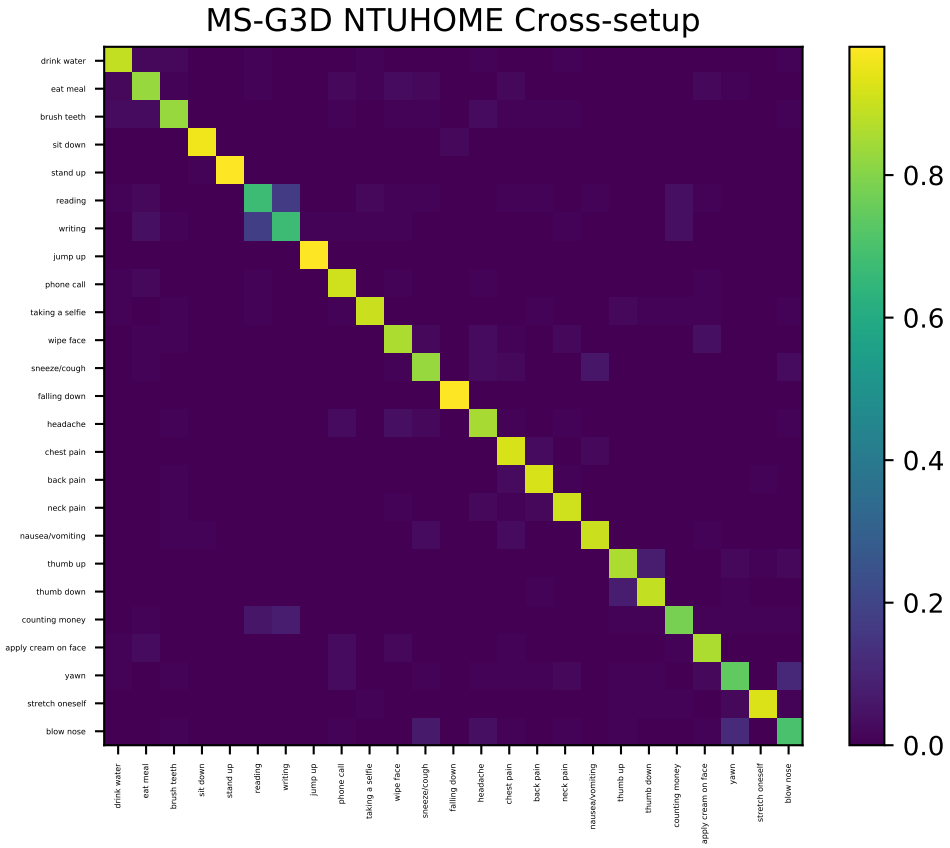
derivados do MS-G3D e Shift-GCN, além das informações de arestas apresentarem o melhor desempenho para ambos.

Tabela 8 – Resultados dos reconhecedores pelo método de validação *cross-setup* na base NTU RGB+D HOME.

	MS-G3D		Shift-GCN	
Dados	**Top 1**	**Top 5**	**Top 1**	**Top 5**
Joint	83,56%	96,98%	80,86%	95,97%
Bone	83,75%	97,05%	81,47%	96,71%
Joint Motion	-	-	79,92%	95,88%
Bone Motion	-	-	81,16%	96,76%
Fusão	86,08%	97,74%	85,95%	98,03%

Fonte: Autor.

Figura 42 – Matriz de confusão obtida pelo MS-G3D na validação *cross-setup* na base NTU RGB+D HOME.

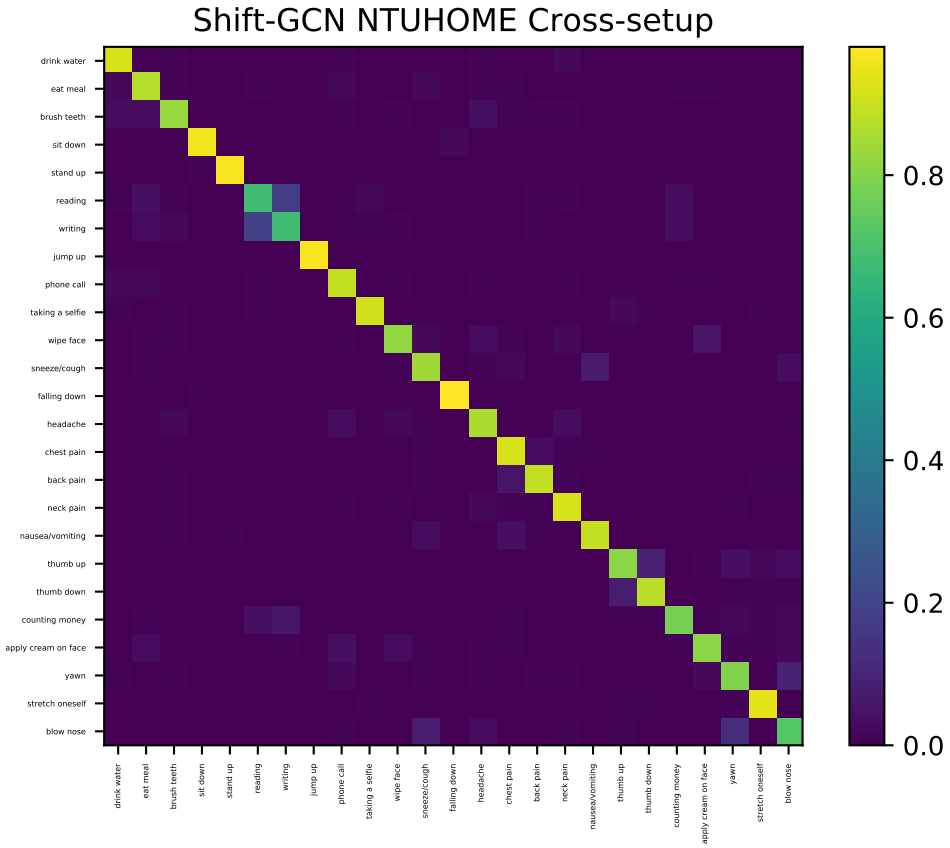


Fonte: Autor.

Os desempenhos obtidos para as precisões Top 5 também reforçam a robustez do reconhecedores para *setups* diferentes, uma vez que eles se encontram entre 95,88% e 97,05%.

É possível observar que a fusão das classificações dos modelos resultou em um aumento de desempenho, sendo estes de 2,33% e 4,48% para os reconhecedores derivados do MS-G3D e Shift-GCN, respectivamente. As matrizes de confusão desse procedimento se encontram ilustradas pelas Figuras 42 e 43 e é possível observar novamente a dificuldade do reconhecedor em distinguir ações compostas por movimentações semelhantes tais como ler e escrever.

Figura 43 – Matriz de confusão obtida pelo Shift-GCN na validação *cross-setup* na base NTU RGB+D HOME.



Fonte: Autor.

6.3 DISCUSSÃO

Dado os resultados dos experimentos apresentados nas Seções anteriores, esta Seção busca discutir e avaliá-los. Assim, características tais como os desempenhos obtidos, volumes das bases de dados utilizadas, tipo de dados mais informativos e possíveis expansões futuras dos reconhecedores de ações baseados em esqueletos 3D serão discutidos.

Os reconhecedores avaliados mostraram desempenhos estado da arte e de acordo com suas propostas originais, além de demonstrar robustez para variações de voluntários e *setups*. Por outro lado, é possível verificar que, em geral, os reconhecedores apresentaram um desempenho maior no método de validação *cross-setup*, sendo esse um indício de que métodos baseados em esqueletos 3D sofrem pouca influência da mudança de ambiente e perspectiva assim como é destacado em Ziyu Liu et al. (2020).

Além disso, pode-se observar pelas matrizes de confusão que os reconhecedores apresentaram dificuldades para diferenciar ações tais como ler e escrever ou dedão para cima e dedão para baixo. Isso indica uma dificuldade em distinguir ações que se diferenciam por movimentações de pequenas partes do corpo como nos casos descritos anteriormente, onde o voluntários realizam os mesmos movimentos no corpo, porém com diferentes movimentações de suas mãos.

É possível observar que as informações de arestas e movimentações de juntas mostraram um melhor desempenho de reconhecimento. Este comportamento pode ser sido obtido pois, enquanto as informações de juntas apresentam somente o posicionamento das partes do esqueleto em um eixo 3D, as arestas e movimentações representam o ângulo e direção da movimentação, sendo isso um indício de que as informações de movimentação e orientação são mais informativas para o reconhecimento de ações do que as posições absolutas das juntas do esqueleto 3D. Além disso, os resultados mostram que a fusão de reconhecedores treinados com diferentes tipos de informações pode aumentar o desempenho da classificação de ações significativamente.

Outra característica importante pode ser observada pelos desempenhos obtidos nas duas bases de dados. Enquanto a NTU RGB+D 120 é a mais volumosa com 114480 amostras, a base NTU RGB+D HOME tem 23784 e, apesar da diferença de quantidade de amostras, os desempenhos obtidos foram semelhantes para ambas as bases, indicando que os reconhecedores não sofreram perda de desempenho pelo menor do volume de dados. Além disso, uma possível adição para trabalhos futuros é a utilização de reconhecedores treinados previamente em bases volumosas, tais como as bases Kinetics e NTU RGB+D 120, para melhorar o desempenho dos reconhecedores na base de dados NTU RGB+D HOME.

Por último, assim como descrito em Ziyu Liu et al. (2020), os métodos de reconhecimento de ações baseados em esqueletos 3D mostram desempenhos estado da arte na literatura recente, porém a extração do esqueleto não permite a analise de informações espaciais tais como objetos utilizados nos *frames* ou o ambiente aonde a ações é executada. Portanto, uma área ainda pouco explorada na literatura recente é utilização de métodos baseados em GCN de

forma multimodal como, por exemplo, a adição de uma RNC que avalia informações espaciais em conjunto com as movimentações que são analisadas pela GCN.

7 CONCLUSÃO

O trabalho proposto apresenta um estudo sobre os métodos de reconhecimento de comportamento humano no nível semântico de ações. Uma revisão bibliográfica foi executada com o objetivo de identificar o estado da arte das bases de dados e dos reconhecedores de ações humanas. Assim, foi possível verificar que existem 2 modelos de bases de dados predominantes na literatura, onde um é a coleta de amostras por meio de plataformas online como o Youtube e outro é a coleta de suas amostras por meio de voluntários em ambientes controlados, onde o último contém rótulos com mais informações além da ação sendo realizada tais como perspectiva e ambiente de coleta.

Dois métodos recentes na literatura foram escolhidos para sua comparação e treinamento nas bases de dados NTU RGB+D 120 e NTU RGB+D HOME. Os métodos escolhidos foram o MS-G3D e o Shift-GCN e reconhecedores foram gerados a partir deles, que foram validados em ambas as bases pelos métodos do *cross-subject* e *cross-setup* de acordo com o protocolo proposto em Jun Liu et al. (2019). Os reconhecedores obtidos apresentaram resultados estado da arte, onde, para a base de dados NTU RGB+D 120, aqueles derivados do MS-G3D mostraram um desempenho de até 84,36% e 85,80% e os derivados do Shift-GCN apresentaram até 83,42% e 85,13% de desempenho para os métodos *cross-subject* e *cross-setup*, respectivamente.

Desempenhos semelhantes foram obtidos na base de dados NTU RGB+D HOME, onde observa-se precisões de até 84,32% e 83,75% para os reconhecedores derivados do MS-G3D e 81,17% e 81,47% para aqueles derivados do Shift-GCN pelos métodos de validação *cross-subject* e *cross-setup*, respectivamente. Assim, é possível verificar que os reconhecedores podem ser treinados em uma base menos volumosa e com uma quantidade menor de ações.

É possível observar que os reconhecedores baseados em esqueletos 3D têm uma maior robustez em relação a *setups* diferentes do que em relação a voluntários diferentes, porém eles não são capazes de analisar informações espaciais como ambientes e objetos diferentes. Uma vez que há poucos trabalhos na literatura que exploram a utilização de dados multimodais no processo de reconhecimento de ações, acredita-se que a adição de um método capaz de reconhecer relações espaciais como uma RNC pode ser aplicado em conjunto com os métodos utilizados neste trabalho para melhorar o desempenho dos reconhecedores. Também é possível verificar que a fusão de reconhecedores que trabalham com base em informações diferentes, como juntas e arestas, podem aumentar o desempenho da classificação de ações de forma considerável

Por último, apesar de existir bases de dados volumosas na literatura recente, elas não disponibilizam uma grande variedade dos voluntários em características tais como idade, nacionalidade e altura. Assim, trabalhos futuros podem envolver a coleta de bases de dados multimodais e com foco na diversidade das características desses voluntários.

REFERÊNCIAS

AGARWAL, I.; KUSHWAHA, A. K. S.; SRIVASTAVA, R. Weighted Fast Dynamic Time Warping based multi-view human activity recognition using a RGB-D sensor. In: 2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG). [S.l.: s.n.], dez. 2015. P. 1–4. DOI: 10.1109/NCVPRIPG.2015.7490046.

AQUINO JUNIOR, Plinio et al. HERA: Home Environment Robot Assistant. In:

BINGBING NI et al. Integrating multi-stage depth-induced contextual information for human action recognition and localization. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). [S.l.: s.n.], abr. 2013. P. 1–8. DOI: 10.1109/FG.2013.6553756.

BLANK, M. et al. Actions as space-time shapes. In: TENTH IEEE International Conference on Computer Vision (ICCV'05) Volume 1. [S.l.: s.n.], 2005. v. 2, 1395–1402 vol. 2. DOI: 10.1109/ICCV.2005.28.

BRUNA, Joan et al. Spectral Networks and Locally Connected Networks on Graphs. **CoRR**, abs/1312.6203, 2013.

CHEN, Chen; LIU, Kui; KEHTARNAVAZ, Nasser. Real-time human action recognition based on depth motion maps. **Journal of Real-Time Image Processing**, v. 12, p. 155–163, 2013.

CHENG, Ke et al. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2020.

CHOUTAS, Vasileios et al. PoTion: Pose MoTion Representation for Action Recognition. **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**, p. 7024–7033, 2018.

DEFFERRARD, Michaël; BRESSON, Xavier; VANDERGHEYNST, Pierre. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In: NIPS. [S.l.: s.n.], 2016.

DOBHAL, Tushar et al. Human Activity Recognition using Binary Motion Image and Deep Learning. **Procedia Computer Science**, v. 58, p. 178–185, 2015. Second International Symposium on Computer Vision and the Internet (VisionNet'15). ISSN 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.08.050. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050915021614>.

DONAHUE, Jeff et al. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 39, p. 677–691, 2014.

DUVENAUD, David et al. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In: NIPS. [S.l.: s.n.], 2015.

FEICHTENHOFER, Christoph; PINZ, Axel; ZISSERMAN, Andrew. Convolutional Two-Stream Network Fusion for Video Action Recognition. **CoRR**, abs/1604.06573, 2016. arXiv: 1604.06573. Disponível em: <http://arxiv.org/abs/1604.06573>.

FOGGIA, P. et al. Exploiting the deep learning paradigm for recognizing human actions. In: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). [S.l.: s.n.], ago. 2014. P. 93–98. DOI: 10.1109/AVSS.2014.6918650.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. [S.l.]: MIT Press, 2016. http://www.deeplearningbook.org.

GORI, Michele; MONFARDINI, Gabriele; SCARSELLI, Franco. A new model for learning in graph domains. **Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.**, v. 2, 729–734 vol. 2, 2005.

HARA, Kensho; KATAOKA, Hirokatsu; SATOH, Yutaka. Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition. **2017 IEEE International Conference on Computer Vision Workshops (ICCVW)**, p. 3154–3160, 2017.

HAYKIN, Simon S. **Neural networks and learning machines**. Third. Upper Saddle River, NJ: Pearson Education, 2009.

HOU, Yonghong et al. Skeleton Optical Spectra-Based Action Recognition Using Convolutional Neural Networks. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 28, p. 807–811, 2018.

IJJINA, E. P.; MOHAN, C. K. Human Action Recognition Based on MOCAP Information Using Convolution Neural Networks. In: 2014 13th International Conference on Machine Learning and Applications. [S.l.: s.n.], dez. 2014. P. 159–164. DOI: 10.1109/ICMLA.2014.30.

JI, S. et al. 3D Convolutional Neural Networks for Human Action Recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 35, n. 1, p. 221–231, jan. 2013. DOI: 10.1109/TPAMI.2012.59.

JIAO, Jichao et al. A Post-Rectification Approach of Depth Images of Kinect v2 for 3D Reconstruction of Indoor Scenes. **ISPRS International Journal of Geo-Information**, v. 6, p. 349, nov. 2017. DOI: 10.3390/ijgi6110349.

KAY, Will et al. **The Kinetics Human Action Video Dataset**. [S.l.: s.n.], 2017. arXiv: 1705.06950 [cs.CV].

KE, Qiuhong et al. Learning Clip Representations for Skeleton-Based 3D Action Recognition. **IEEE Transactions on Image Processing**, v. 27, p. 2842–2855, 2018.

KHAN, F. M.; SINGH, V. K.; NEVATIA, R. Simultaneous inference of activity, pose and object. In: 2012 IEEE Workshop on the Applications of Computer Vision (WACV). [S.l.: s.n.], jan. 2012. P. 281–288. DOI: 10.1109/WACV.2012.6163003.

KIPF, Thomas; WELLING, Max. Semi-Supervised Classification with Graph Convolutional Networks. **ArXiv**, abs/1609.02907, 2016.

KONDA, Kishore et al. Real-time activity recognition via deep learning of motion features. In: ESANN. [S.l.: s.n.], 2015.

KUEHNE, Hilde et al. HMDB: A large video database for human motion recognition. **2011 International Conference on Computer Vision**, p. 2556–2563, 2011.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, nov. 1998. ISSN 0018-9219. DOI: 10.1109/5.726791.

LEE, Inwoong et al. Ensemble Deep Learning for Skeleton-Based Action Recognition Using Temporal Sliding LSTM Networks. **2017 IEEE International Conference on Computer Vision (ICCV)**, p. 1012–1020, 2017.

LI, Chao et al. Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation. In: IJCAI. [S.l.: s.n.], 2018.

LI, Chaolong et al. Spatio-Temporal Graph Convolution for Skeleton Based Action Recognition. In: AAAI. [S.l.: s.n.], 2018.

LI, Dong et al. Unified Spatio-Temporal Attention Networks for Action Recognition in Videos. **IEEE Transactions on Multimedia**, v. 21, p. 416–428, 2019.

LI, W.; ZHANG, Z.; LIU, Z. Action recognition based on a bag of 3D points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. [S.l.: s.n.], jun. 2010. P. 9–14. DOI: 10.1109/CVPRW.2010.5543273.

LI, Yingwei; LI, Yi; VASCONCELOS, Nuno. RESOUND: Towards Action Recognition Without Representation Bias. In: ECCV. [S.l.: s.n.], 2018.

LI, Yujia et al. Gated Graph Sequence Neural Networks. **CoRR**, abs/1511.05493, 2016.

LIU, Jun et al. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. **IEEE transactions on pattern analysis and machine intelligence**, 2019.

LIU, Jun et al. Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks. **IEEE Transactions on Image Processing**, v. 27, p. 1586–1599, 2018.

LIU, Kun et al. T-C3D: Temporal Convolutional 3D Network for Real-Time Action Recognition. In: AAAI. [S.l.: s.n.], 2018.

LIU, Ziyu et al. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. **2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, p. 140–149, 2020.

LUVIZON, Diogo C.; PICARD, David; TABIA, Hedi. 2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning. **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**, p. 5137–5146, 2018.

MEDHEKAR, Aruna et al. **Accelerate OpenCV: Optical Flow Algorithms with NVIDIA Turing GPUs**. [S.l.: s.n.], dez. 2019. Disponível em: <https://developer.nvidia.com/blog/opencv-optical-flow-algorithms-with-nvidia-turing-gpus/>.

MESHRY, M.; HUSSEIN, M. E.; TORKI, M. Linear-time online action detection from 3D skeletal data using bags of gesturelets. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). [S.l.: s.n.], mar. 2016. P. 1–9. DOI: 10.1109/WACV.2016.7477587.

OFLI, F. et al. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV). [S.l.: s.n.], 2013. P. 53–60. DOI: 10.1109/WACV.2013.6474999.

PAPAKOSTAS, M. et al. Short-Term Recognition of Human Activities Using Convolutional Neural Networks. In: 2016 12th International Conference on Signal-Image Technology Internet-Based Systems (SITIS). [S.l.: s.n.], nov. 2016. P. 302–307. DOI: 10.1109/SITIS.2016.56.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. **Nature**, v. 323, p. 533–536, 1986.

SCARSELLI, Franco et al. The Graph Neural Network Model. **IEEE Transactions on Neural Networks**, v. 20, p. 61–80, 2009.

SHAHROUDY, Amir et al. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, p. 1010–1019, 2016.

SI, Chenyang et al. Skeleton-Based Action Recognition with Spatial Reasoning and Temporal Stack Learning. **ArXiv**, abs/1805.02335, 2018.

SIMONYAN, Karen; ZISSERMAN, Andrew. Two-Stream Convolutional Networks for Action Recognition in Videos. **ArXiv**, abs/1406.2199, 2014.

SMAIRA, Lucas et al. **A Short Note on the Kinetics-700-2020 Human Action Dataset**. [S.l.: s.n.], 2020. arXiv: 2010.10864 [cs.CV].

SONG, Sijie et al. Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection. **IEEE Transactions on Image Processing**, v. 27, p. 3459–3471, 2018.

SOOMRO, Khurram; ZAMIR, Amir Roshan; SHAH, Mubarak. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. **ArXiv**, abs/1212.0402, 2012.

TANFOUS, Amor Ben; DRIRA, Hassen; AMOR, Boulbaba Ben. Coding Kendall's Shape Trajectories for 3D Action Recognition. **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**, p. 2840–2849, 2018.

TU, Zhigang et al. Multi-stream CNN: Learning representations based on human-related regions for action recognition. **Pattern Recognition**, v. 79, p. 32–43, 2018.

ULLAH, Amin et al. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. **IEEE Access**, v. 6, p. 1155–1166, 2018.

VAQUETTE, G. et al. The DAily Home LIfe Activity Dataset: A High Semantic Activity Dataset for Online Recognition. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017). [S.l.: s.n.], mai. 2017. P. 497–504. DOI: 10.1109/FG.2017.67.

VEMULAPALLI, R.; ARRATE, F.; CHELLAPPA, R. Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], jun. 2014. P. 588–595. DOI: 10.1109/CVPR.2014.82.

WANG, H.; SCHMID, C. Action Recognition with Improved Trajectories. In: 2013 IEEE International Conference on Computer Vision. [S.l.: s.n.], dez. 2013. P. 3551–3558. DOI: 10.1109/ICCV.2013.441.

WANG, J. et al. Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], jun. 2012. P. 1290–1297. DOI: 10.1109/CVPR.2012.6247813.

WANG, Lei; HUYNH, Du Q.; KONIUSZ, Piotr. A Comparative Review of Recent Kinect-Based Action Recognition Algorithms. **IEEE Transactions on Image Processing**, v. 29, p. 15–28, 2019.

WANG, P. et al. Action Recognition From Depth Maps Using Deep Convolutional Neural Networks. **IEEE Transactions on Human-Machine Systems**, v. 46, n. 4, p. 498–509, ago. 2016. DOI: 10.1109/THMS.2015.2504550.

WANG, Pichao et al. Action recognition based on joint trajectory maps with convolutional neural networks. **ArXiv**, abs/1612.09401, 2018.

WANG, Pichao et al. Deep Convolutional Neural Networks for Action Recognition Using Depth Map Sequences. **CoRR**, abs/1501.04686, 2015. arXiv: 1501.04686. Disponível em: <http://arxiv.org/abs/1501.04686>.

WANG, Pichao et al. Depth Pooling Based Large-Scale 3-D Action Recognition With Convolutional Neural Networks. **IEEE Transactions on Multimedia**, v. 20, p. 1051–1061, 2018.

WANG, Xuanhan et al. Two-Stream 3-D convNet Fusion for Action Recognition in Videos With Arbitrary Size and Length. **IEEE Transactions on Multimedia**, v. 20, p. 634–644, 2018.

WU, Zonghan et al. A Comprehensive Survey on Graph Neural Networks. **ArXiv**, abs/1901.00596, 2019.

XIA, L.; CHEN, C.C.; AGGARWAL, JK. View invariant human action recognition using histograms of 3D joints. In: IEEE. COMPUTER Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. [S.l.: s.n.], 2012. P. 20–27.

YAN, Sijie; XIONG, Yuanjun; LIN, Dahua. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In: AAAI. [S.l.: s.n.], 2018.

YONG DU; WANG, W.; WANG, L. Hierarchical recurrent neural network for skeleton based action recognition. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], jun. 2015. P. 1110–1118. DOI: 10.1109/CVPR.2015.7298714.

YU, Bing; YIN, Haoteng; ZHU, Zhanxing. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In: IJCAI. [S.l.: s.n.], 2018.

ZHANG, Hong-Bo et al. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. In: SENSORS. [S.l.: s.n.], 2019.

ZHANG, Jing et al. RGB-D-based action recognition datasets: A survey. **Pattern Recognition**, v. 60, p. 86–105, 2016. ISSN 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2016.05.019. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320316301029>.

ZHU, Yi et al. Hidden Two-Stream Convolutional Networks for Action Recognition. In: ACCV. [S.l.: s.n.], 2017.