UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus Bauru

Bruno Belluzzo

Reconhecimento de Ações Humanas Baseado em Articulações do Esqueleto Obtidas de Poses 2D

Bauru
2023

Bruno Belluzzo

Reconhecimento de Ações Humanas Baseado em Articulações do Esqueleto Obtidas de Poses 2D

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de Bauru.

Orientador: Prof. Associado Aparecido Nilceu Marana

Bauru
2023

Belluzzo, Bruno.
Reconhecimento de Ações Humanas Baseado em Articulações do Esqueleto Obtidas de Poses 2D / Bruno Belluzzo. – Bauru, 2023
69 f. : il., tabs.

Orientador: Prof. Associado Aparecido Nilceu Marana
Dissertação (mestrado) - Universidade Estadual Paulista “Júlio de Mesquita Filho", Instituto de Biociências, Letras e Ciências Exatas

1. Reconhecimento de ações humanas. 2. Estimação de pose humana. 3. Aprendizado em Profundidade. I. Marana, Aparecido Nilceu II. Universidade Estadual Paulista “Júlio de Mesquita Filho", Instituto de Biociências, Letras e Ciências Exatas. III. Reconhecimento de Ações Humanas Baseado em Articulações do Esqueleto Obtidas de Poses 2D.

CDU – 518.72:76

Ficha catalográfica elaborada pela biblioteca do IBILCE
UNESP - Câmpus de São José do Rio Preto

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Bauru

Faculdade de Ciências

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE BRUNO BELLUZZO, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 03 dias do mês de fevereiro do ano de 2023, às 09:00 horas, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de BRUNO BELLUZZO, intitulada **Reconhecimento de Ações Humanas Baseado em Articulações do Esqueleto Obtidas de Poses 2D**. A Comissão Examinadora foi constituida pelos seguintes membros: Prof. Dr. APARECIDO NILCEU MARANA (Orientador(a) - Participação Virtual) do(a) Departamento de Computacao / Faculdade de Ciências - UNESP - Bauru, Prof. Dr. KELTON AUGUSTO PONTARA DA COSTA (Participação Virtual) do(a) Departamento de Computacao / Faculdade de Ciências - UNESP - Bauru, Prof. Dr. MURILO VARGES DA SILVA (Participação Virtual) do(a) Instituto Federal de São Paulo. Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final: Aprovado _ _ _ _ _ _ . Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Prof. Dr. APARECIDO NILCEU MARANA

Este trabalho é dedicado ao meu pai e à minha mãe.

Agradecimentos

Gostaria de agradecer primeiramente aos meus pais, Neto (in memoriam) e Rosangela, os quais sempre me apoiaram em todas as minhas decisões e não mediram esforços para que eu conseguisse realizar todos os meus sonhos e anseios profissionais.

Aos meus irmãos, Giovani e Pietro, pelos momentos de descontração e por estarem sempre ao meu lado, nos bons e nos maus momentos, sempre me apoiando a seguir em frente.

À minha namorada e futura esposa, Luiza, que esteve ao meu lado praticamente em toda minha vida, mas que me acompanha lado a lado desde 2014, tendo vivido meus altos e baixos por todo esse tempo. Agradeço a paciência, carinho e motivação.

À minha família e amigos, por todos os momentos em que me encorajaram a seguir em frente.

Ao professor doutor Aparecido Nilceu Marana, o qual tenho profundo carinho por ter me orientado durante todos esses anos durante o mestrado, não existem palavras para expressar minha gratidão pela paciência, motivação e ensinamentos que compartilhamos durante essa caminhada, nada disso teria sido possível sem você.

A persistência é o caminho do êxito.
Charles Chaplin

Resumo

Com o aumento da capacidade das tecnologias atuais de armazenamento e processamento de grandes volumes de dados em uma velocidade cada vez maior, a análise e o reconhecimento de padrões em vídeos passaram a ser pesquisadas e empregadas nas mais diversas aplicações, dentre as quais o reconhecimento automático de ações humanas, que visa identificar em um determinado vídeo as ações sendo executadas pelas pessoas presentes, seja para fins recreativos ou para o monitoramento e a segurança em locais públicos ou até mesmo privados. Detectar pessoas nos vídeos e reconhecer as ações sendo realizadas por elas é uma tarefa complexa, pois exige a extração de caracteríticas que representam um padrão de movimentos realizados pela pessoa tanto no aspecto espacial, quanto no aspecto temporal, ao longo dos diversos *frames* do vídeo. Uma maneira de obter informações que descrevam o movimento do corpo humano em vídeos é identificar as articulações do esqueleto humano nos diversos *frames*, o que pode ser realizado utilizando-se algoritmos de estimação de pose 2D em imagens. Atualmente, existem algoritmos bastante eficazes e eficientes disponíveis, capazes de detectar as articulações do corpo humano e retornarem suas coordenadas nas imagens. Aliado a isso, tem se observado nos últimos anos uma grande evolução dos métodos e algoritmos de aprendizado de máquina, destinados ao reconhecimento de padrões complexos, inspirados em modelos biológicos, com ênfase nos métodos baseados em aprendizado de máquina profundo e recorrente. Esta dissertação de mestrado tem como objetivo propor um método de reconhecimento de ações humanas em vídeo baseado nas articulações dos esqueletos obtidas de poses 2D estimadas por meio de algoritmos estado da arte, utilizando redes neurais recorrentes convolucionais para propiciar mais robustez ao processo. O método proposto foi avaliado utilizando-se duas bases de dados públicas e populares de vídeos de ações humanas, a KTH e a Weizmann. Os resultados obtidos foram superiores aos resultados obtidos por vários métodos encontrados na literatura e comparáveis à métodos estado-da-arte, com a vantagem de apresentar uma estratégia simples para a extração de características a partir das articulações dos esqueletos obtidas das poses 2D.

Palavras-chave: *Reconhecimento de Ações Humanas, Poses 2D, Aprendizado em Profundidade, Redes Neurais Recorrentes.*

Abstract

With the increase in the capacity of current technologies for storing and processing large volumes of data at an ever-increasing speed, the analysis and recognition of patterns in videos began to be researched and used in the most diverse applications, among which automatic recognition of human actions, which aims to identify in a given video the actions being performed by the people present, whether for recreational purposes or for monitoring and security in public or even private places. Detecting people in the videos and recognizing their actions is a complex task, as it requires the extraction of features that represent a pattern of movements performed by the person both in the spatial and temporal aspects, along the different frames of the video. One way to obtain information describing the movement of the human body in videos is to identify the joints of the human skeleton in the different frames, which can be done using 2D pose estimation algorithms in images. Currently, there are very effective and efficient algorithms available, capable of detecting the joints of the human body and returning their coordinates in the images. Allied with this, there has been a great evolution in machine learning methods and algorithms in recent years, aimed at recognizing complex patterns, inspired by biological models, with emphasis on methods based on deep and recurrent machine learning. This master's thesis aimed to propose a method for recognizing human actions in video based on skeletal joints obtained from 2D poses estimated using state-of-the-art algorithms, using deep machine learning methods and recurrent neural networks to provide more robustness to the process. The proposed method was evaluated using two public and popular databases of videos of human actions, KTH and Weizmann. The results obtained were superior to several methods found in the literature and comparable to state-of-the-art methods, with the advantage of presenting a simple strategy for extracting features from skeletal joints obtained from 2D poses.

Keywords: *Human Action Recognition, 2D Poses, Deep Learning, Recurrent Neural Networks.*

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ASD	Accumulated Squared Derivative
BoVW	Bag-of-Visual-Words
SJoBB	Skeleton Joints on Black Background
CCD	Charge-Coupled Device
CNN	Convolutional Neural Network
CPU	Central Process Unit
DT	Dense Trajectories
DTW	Dynamic Time Warping
ELM	Extreme Learning Machine
FPS	Frames Por Segundo
GLCM	Gray-Level Co-occurrence Matrix
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
HOF	Histograms of Optical Flow
HOG	Histograms of Oriented Gradients
HOT	Histogram of Oriented Trajectories
iDT	Improved Dense Trajectories
LSTM	Long Short-Term Memory
MBH	Motion Boundary Histograms
NN	Nearest Neighbor
PAF	Part Association Field
PAFs	Part Association Fields
PCA	Principal Component Analysis

PIF	Part Intensity Field
RAH	Reconhecimento de Ações Humanas
ResNet	Residual Network
RNC	Redes Neurais Convolucionais
RNN	Recurrent Neural Network
SAX	Symbolic Aggregate Approximation
SVM	Support Vector Machine
VR	Visual Rhythm

Sumário

1 Introdução

O Reconhecimento de Ações Humanas (RAH) é uma área de pesquisa que visa entender o comportamento humano e atribuir um rótulo a cada ação. O RAH possui uma ampla gama de aplicações e, por isso, vem atraindo cada vez mais atenção de vários pesquisadores e empresas. As ações humanas podem ser representadas usando várias modalidades de dados, como imagens RGB, esqueleto, profundidade, infravermelho, nuvem de pontos, fluxo de eventos, áudio, aceleração, radar e sinal WiFi, que codificam diferentes fontes de informações úteis, mas distintas, e têm várias vantagens, dependendo nos cenários de aplicação (SUN et al., 2022).

Algumas aplicações importantes do RAH no mundo real envolvem monitoramento em áreas públicas ou privadas, identificando as ações que as pessoas possam estar realizando. Por exemplo, no monitorando automático de pessoas idosas, os familiares podem ser avisados caso os sensores detectem algum evento perigoso, como uma queda, ou algum gesto específico feito pelo idoso sinalizando que está precisando de auxílio (VIJAYAPRABAKARAN; SATHIYAMURTHY; PONNIAMMA, 2020). Em esportes, as comissões técnicas podem monitorar e analisar os desempenhos das equipes registrando as ações realizadas pelos seus atletas no decorrer das partidas e levantando suas métricas de performance (RANGASAMY et al., 2020). Na indústria, os gestores podem monitorar os locais de trabalho para verificar se os colaboradores estão exercendo as atividades de forma adequada e segura (PARSA, 2020).

A popularização e o barateamento das câmeras de captura de vídeos, aliados ao aumento da capacidade das tecnologias atuais de armazenamento, processamento e transmissão de grandes volumes de dados em uma velocidade cada vez maior, têm contribuído para aumentar a demanda por sistemas computacionais destinados à aplicações baseadas em análise e reconhecimento automático de padrões em vídeos. Dentre as aplicações mais relevantes que requerem o processamento de vídeos encontram-se aquelas que visam o reconhecimento de ações humanas (RAH) (PAREEK; THAKKAR, 2021).

Dentre os principais desafios da tarefa de RAH baseado em vídeos, destacam-se: as oclusões parciais do corpo das pessoas, a presença de movimentos irregulares, a diversidade dos cenários nos quais as ações são realizadas que gera uma grande heterogeneidade no fundo das imagens, os diferentes ângulos de visão das câmeras utilizadas, etc (WANG; SCHMID, 2013). Esses desafios fazem com que os algoritmos de visão computacional apresentem desempenhos muito abaixo do desempenho humano em tarefas de RAH. Muitos algoritmos e métodos apresentam bons resultados quando aplicados em bases de dados criadas em laborátorios com ambientes controlados, porém, falham ao fazer o reconhecimentos das ações em ambientes realistas, em diferentes cenários e em situações do cotidiano das pessoas (JHUANG et al., 2013).

Detectar pessoas nos vídeos e reconhecer as ações que elas estão realizando é uma tarefa complexa, pois exige a extração de caracteríticas que representam um padrão de movimentos realizados pela pessoa tanto no aspecto espacial, quanto no aspecto temporal, ao longo dos diversos *frames* do vídeo. Características extraídas de espaço-tempo capturam formas e movimentos peculiares nos vídeos e fornecem representações independentes de escalas e outras complicações nas imagens como alterações de fundo. Alguns descritores de características buscam capturar a forma e o movimento nas vizinhanças de pontos selecionados usando medições de imagem como gradientes de imagem espaciais ou espaço-temporais e fluxo óptico (WANG et al., 2009).

Recentemente, com o surgimento de novas técnicas de estimação de pose humana 2D, como o OpenPose (CAO et al., 2019) e o PifPaf (KREISS; BERTONI; ALAHI, 2019), caracteríticas provenientes dos esqueletos gerados por essas tecnologias passaram a ser utilizadas para o reconhecimento de ações humanas, uma vez que elas podem abstrair informações do fundo das imagens e preservar a privacidade das pessoas presentes nos vídeos (SILVA, 2020).

1.1 Objetivos

Este projeto de dissertação de mestrado visa propor e desenvolver métodos para o reconhecimento de ações humanas em vídeos baseados em métodos de estimação de pose 2D, utilizando técnicas de aprendizado de máquina em profundidade, em particular as Redes Neurais Convolucionais (CNN - Convolutional Neural Networks), para a análise espacial das ações, e as Redes Neurais Recorrentes (RNN - Recurrent Neural Networks), em particular as redes neurais Long Short Term Memory (LSTM).

A principal hipótese deste trabalho é que o uso das informações obtidas das poses humanas 2D associadas com métodos adequados de aprendizado de máquina podem contribuir para avançar o estado da arte de métodos de reconhecimento automático de ações humanas em vídeos.

1.2 Organização da Dissertação

Esta dissertação está organizada em seis capítulos, a saber:

Capítulo 1 : Apresenta a introdução ao tema da dissertação de mestrado e seus objetivos;

Capítulo 2: Apresenta a fundamentação teórica do trabalho, descrevendo os conceitos de estimação de pose humana, redes neurais convolucionais, redes neurais recorrentes e de mecanismos de atenção;

Capítulo 3: Apresenta os princípios do RAH, introduzindo seus principais conceitos e diferentes categorias de ações a serem classificadas. Apresenta também os métodos correlatos existentes na literatura para reconhecimento de ações humanas e as principais bases de dados utilizadas na literatura nas mais diversas pesquisas;

Capítulo 4: Apresenta o método proposto para o reconhecimento de ações humanas baseado em articulações do esqueleto, obtidas de poses 2D;

Capítulo 5: Apresenta os experimentos realizados para avaliação do método proposto, bem como os resultados obtidos e uma comparação com métodos correlatos encontrados na literatura;

Capítulo 6: Apresenta as conclusões do trabalho, direcionamentos para trabalhos futuros e a publicação realizada.

2 Fundamentação Teórica

Neste capítulo apresenta-se a fundamentação teórica, que inclui o conceito de estimação de pose humana, de redes neurais convolucionais, de redes neurais recorrentes e de mecanismos de atenção.

2.1 Estimação de Pose Humana

O problema de localizar pontos-chave anatômicos ou partes do corpo de indivíduos é conhecido como estimação de pose humana. Inferir a pose de várias pessoas em imagens, apresenta um conjunto de desafios. Primeiro, cada imagem pode conter um número de pessoas desconhecido que podem ocorrer em qualquer posição ou escala. Segundo, as interações entre as pessoas induzem interferências espaciais complexas, devido ao contato, à oclusão e às articulações dos membros, dificultando a associação das partes. Terceiro, o tempo de execução tende a crescer conforme cresce o número de pessoas na imagem, tornando o desempenho em tempo real um desafio (CAO et al., 2019).

Uma abordagem comum consiste em primeiro detectar as pessoas na imagem e em seguida detectar as articulações em cada pessoa detectada. Esse tipo de abordagem, conhecida como *top-down*, potencializa diretamente as técnicas existentes para estimativa de pose de uma pessoa, mas é bastante impactada com os erros nos estágios iniciais: se a detecção da pessoa falhar, o que é propenso a acontecer quando as pessoas estão muito próximas, não há como recuperar o erro. Além disso, o tempo de execução dessas abordagens de cima para baixo é proporcional ao número de pessoas: para cada detecção, é calculada uma estimativa de pose, e quanto mais pessoas houver, maior será o custo computacional.

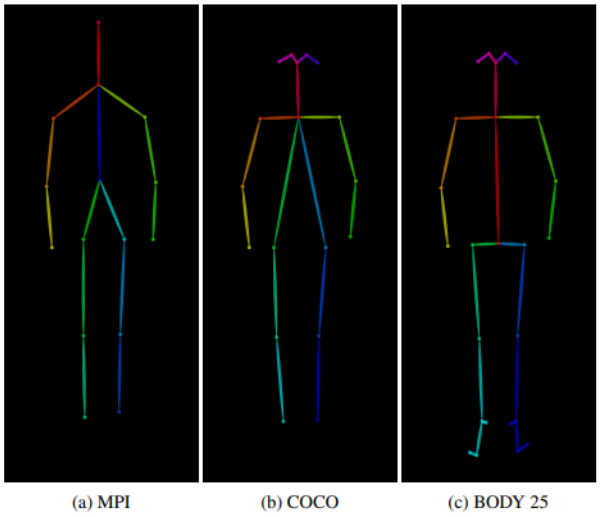
Em contraste, as abordagens *bottom-up* são atraentes uma vez que oferecem robustez às atividades iniciais e tem o benefício de desacoplar a complexidade do tempo de execução do número de pessoas na imagem. No entanto, as abordagens de baixo para cima não usam diretamente dicas contextuais globais de outras partes do corpo e outras pessoas. Na prática, os métodos de baixo para cima antigos não retêm os ganhos de eficiência, pois a análise final requer inferências globais custosas (CAO et al., 2019).

Atualmente, a maioria dos algoritmos de estimativa de pose são baseados em redes neurais, majoritariamente em redes neurais convolucionais, superando os resultados dos métodos tradicionais. Alguns algoritmos com soluções baseadas em redes neurais profundas também atingiram boa performance, é o caso do DeepPose, que alavancou uma série de trabalhos que usam aprendizado profundo para estimativa de pose 2D (TOSHEV; SZEGEDY, 2014).

Com o surgimento de novas bases de dados voltadas para problemas de visão computa-

cional e algumas delas com foco em estimativa de pose, como por exemplo a Microsoft Objects in Context (COCO) (LIN et al., 2014), vários modelos para representar poses 2D surgiram, dentre os quais destacam-se: MPI (INSAFUTDINOV et al., 2016), COCO (LIN et al., 2014) e BODY 25 (CAO et al., 2019). A Figura 1 mostra esses modelos.

Figura 1 – Modelos de pose 2D.



Fonte: (SILVA; MARANA, 2020).

Métodos que utilizam aprendizado profundo em estratégias *bottom-up* para estimar poses 2D com cenas contendo várias pessoas mostraram boa acurácia e tempo de processamento em *real-time*. Dentre esses métodos, dois têm apresentado ótimos resultados para o reconhecimento de ações em vídeos, que são o OpenPose (CAO et al., 2019) e o PifPaf (KREISS; BERTONI; ALAHI, 2019). As Subseções 2.1.1 e 2.1.2 descrevem, respectivamente, estes métodos.

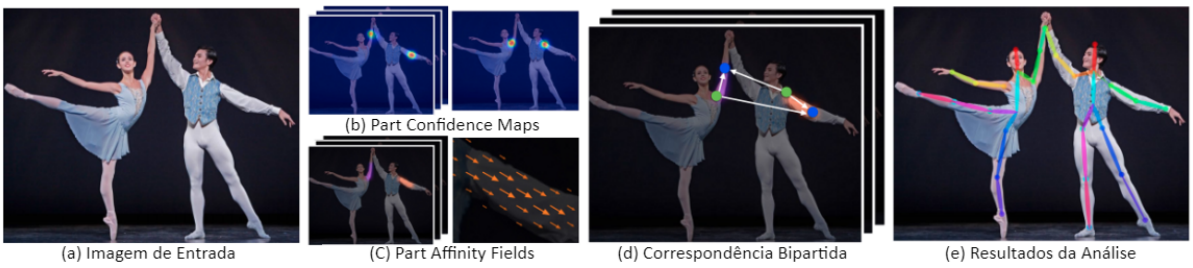
2.1.1 OpenPose

OpenPose é um método eficiente para estimativa de pose 2D em imagens contendo várias pessoas e tem apresentado bom desempenho em várias aplicações distintas. Ele usa uma abordagem *bottom-up* das pontuações de associação via *Part Association Fields* (PAFs), que são um conjunto de campos vetoriais 2D que codificam a localização e orientação dos membros sobre o domínio da imagem.

O *pipeline* do OpenPose segue os seguintes passos: primeiro, uma imagem RGB (Figura 2a) é processada por uma rede neural convolucional que irá produzir duas saídas, uma com um mapa de confiança (Figura 2b) de partes de corpo e outra com os campos de afinidade (Figura 2c), que representa o grau de associação entre as diferentes partes do corpo, em seguida, os mapas de confiança e os campos de afinidade são processados (Figura 2d) e as poses 2D são

estimadas (Figura 2e), incluindo as coordenadas das juntas e a confiança para todas as pessoas detectadas na imagem (M. MAITHANI, 2021).

Figura 2 – Pipeline do método OpenPose.



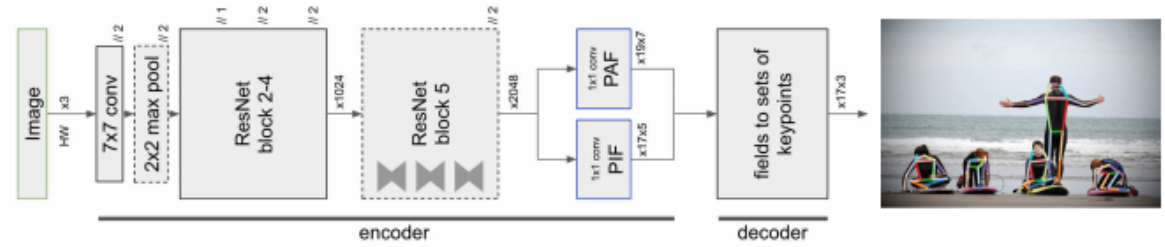
Fonte: Adaptado de (CAO et al., 2019).

2.1.2 PifPaf

O objetivo do método PifPaf é estimar poses humanas em imagens com muitas pessoas. O método tenta lidar com desafios como imagens de baixa resolução e pessoas parcialmente ocluídas. Técnicas que usam a abordagem *top-down* geralmente falham quando as pessoas são obstruídas por outras, quando ocorre a sobreposição de *bounding boxes*. Em geral, os métodos *bottom-up* propostos antes do PifPaf são livres de *bounding-box*, mas eles ainda contêm um mapa de localização de características. No entanto, o método PifPaf é livre de qualquer restrição baseada em grade de localização espacial das articulações e pode detectar poses múltiplas, mesmo quando a oclusão acontece. A principal diferença entre o PifPaf e o OpenPose é que o PifPaf vai além de campos escalares e vetoriais para campos compostos no módulo PAF (KREISS; BERTONI; ALAHI, 2019).

A Figura 3 mostra a arquitetura do PifPaf, que faz uso de uma rede neural residual (ResNet) compartilhada com duas outras redes: uma responsável por prever as juntas do corpo (localização, confiança e tamanho), que é chamada de *Part Intensity Field* (PIF), e outra rede que retorna as associações entre as juntas, chamada *Part Association Field* (PAF), assim dando o nome ao método PifPaf.

Figura 3 – Arquitetura do PifPaf.



Fonte: (KREISS; BERTONI; ALAHI, 2019).

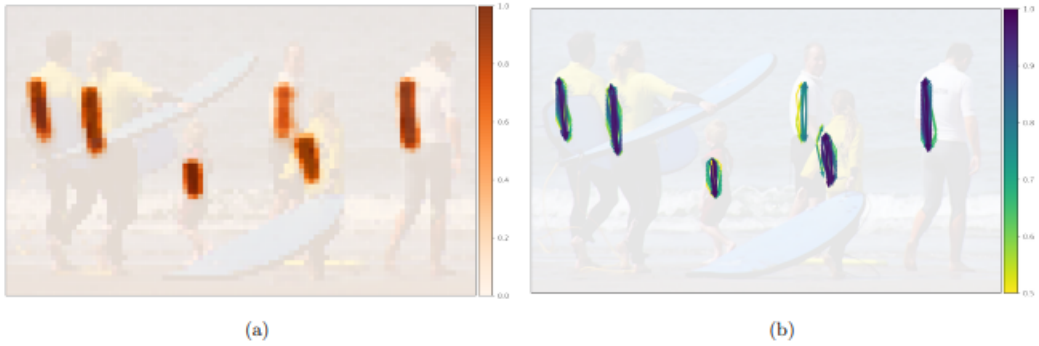
A primeira etapa do método PifPaf é o PIF, que detecta e localiza com precisão as partes do corpo humano. Para fazer a fusão de um mapa de confiança, uma regressão para detecção de pontos-chave é usada. Como resultado do módulo PIF, uma estrutura de junta é gerada incluindo a medida de confiança, um vetor que aponta para a parte do corpo mais próxima e o tamanho da junta. Na Figura 4a, por exemplo, é apresentado o mapa de confiança para os ombros esquerdos presentes na imagem. A fim de melhorar a localização do mapa de confiança, é realizada uma fusão com a parte vetorial do PIF, conforme mostrado na Figura 4b, formando um mapa de confiança de alta resolução (Figura 4c).

Figura 4 – Módulo PIF do método PifPaf.

Fonte: (KREISS; BERTONI; ALAHI, 2019).

O PAF é a etapa para conectar os locais das juntas em poses 2D. Para cada junta, o PAF retorna a localização, a medida de confiança, dois vetores para as duas partes associadas e duas larguras para a precisão espacial. A Figura 5 mostra o PAF que associa o ombro esquerdo com quadril esquerdo (KREISS; BERTONI; ALAHI, 2019).

Figura 5 – Módulo PAF do método PifPaf.



Fonte: (KREISS; BERTONI; ALAHI, 2019).

2.2 Redes Neurais Convolucionais

Com o avanço e disponibilização de poderosos computadores com processamento paralelo, como computação em nuvem, GPUs, *clusters* de CPUs, e o aumento em larga escala

de dados disponíveis para o treinamento de modelos, o interesse no uso de Redes Neurais Convolucionais (CNN, do inglês *Convolutional Neural Network*) cresceu.

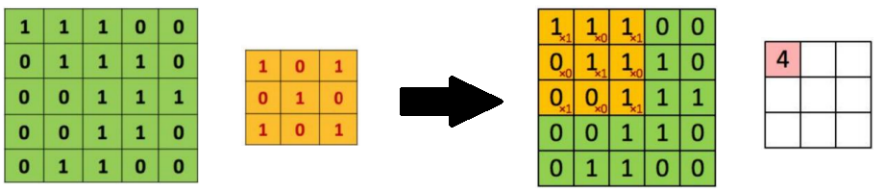
Redes Neurais Convolucionais são um tipo de rede neural artificial que requerem uma camada convolucional, mas também pode conter camadas não lineares, de *pooling* e totalmente conectadas, para criar uma rede neural convolucional profunda. Principalmente problemas que trabalham com imagens podem se beneficiar das camadas de convolução. Na CNN, os filtros convolucionais são treinados usando o método *backpropagation*. As formas da estrutura do filtro dependem da tarefa dada. Por exemplo, em uma aplicação como detecção de rosto, um filtro pode realizar a extração de borda, enquanto outro pode realizar a extração do olho. No entanto, não controlamos totalmente esses filtros na CNN e seus valores são determinados por meio de aprendizado (ALBAWI et al., 2018).

A convolução é uma operação onde um filtro g desliza sobre uma imagem f, ambos de dimensão $m \times n$, aplicando uma função matemática em cada *pixel*, conforme definido na Equação 1, gerando assim novos valores que são utilizados como a saída desta camada de convolução. Dessa forma, acontece uma sumarização dos dados de entrada.

$$(f * g)[m, n] = \sum_{j=-\infty}^{+\infty} \sum_{i=-\infty}^{+\infty} f[i, j] g[m - i, n - j] \tag{1}$$

A Figura 6 mostra um exemplo que ilustra, de forma simplificada, o processo de convolução, no qual, a partir de uma matriz de entrada, de dimensão 5×5 , é aplicada a operação de convolução utilizando-se um filtro de dimensão 3×3, gerando uma matriz de saída cujos elementos são obtidos aplicando-se a Equação 1 (WU, 2017).

Figura 6 – Exemplo da operação de convolução.



Fonte: Elaborada pelo autor.

Nas CNNs, as operações convolucionais são especificadas pelo passo, pelo tamanho do filtro e pelo preenchimento de zeros. O passo determina quantos *pixels* serão deslizados para a direita a cada vez que a convolução ocorrer, para assim se gerar o próximo valor de saída. O preenchimento de zeros adiciona linhas e colunas com valores zero à matriz original para controlar o tamanho do mapa de *features* da saída (O'SHEA; NASH, 2015).

Outro tipo de camada presente em arquiteturas de CNNs são as camadas de *pooling*, responsáveis por diminuir a dimensionalidade dos dados de entrada e conservar apenas as informações importantes de cada região da imagem. Para isso existem operações como *max pooling, average pooling e min pooling*, que a partir de uma região da imagem, a camada de pooling retorna um valor dependendo da operação a ser realizada (LIN; CHEN; YAN, 2013).

A operação de *Max-pooling* recebe uma matriz de entrada $N_{in} \times N_{in}$ e retorna uma matriz de saída menor, digamos $N_{out} \times N_{out}$. Essa matriz de saída é obtida dividindo a de entrada em N_{out}^2 regiões de *pooling* ($P_{i,j}$) (GRAHAM, 2014):

$$P_{i,j} \subset \{1, 2, ..., N_{in}\}^2$$

para cada

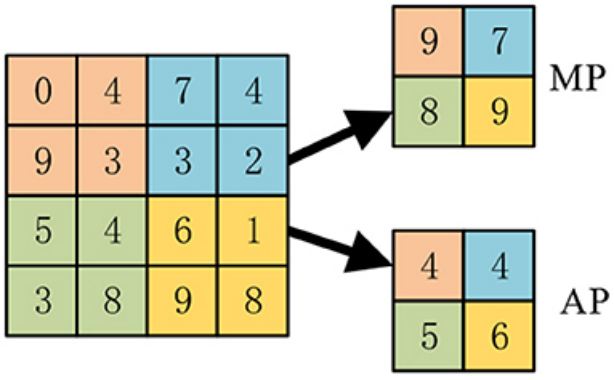
$$(i, j)\, \epsilon\, \{1, ..., N_{out}\}^2$$

obtendo assim

$$Output_{i,j} = \max_{(k,l)\epsilon P_{i,j}} Input_{k,l} \tag{2}$$

A operação de *Average-pooling* é semelhante a de *Max-pooling* mas com a diferença que na Equação 2 se calcula a média dos valores invés do valor máximo. Para exemplificar as equações e ficar mais claro como essa estratégia funciona, a Figura 7 mostra, a partir de uma matriz de entrada, o resultado de saída sendo aplicado uma operação de *max pooling* (MP) e *average pooling* (AP).

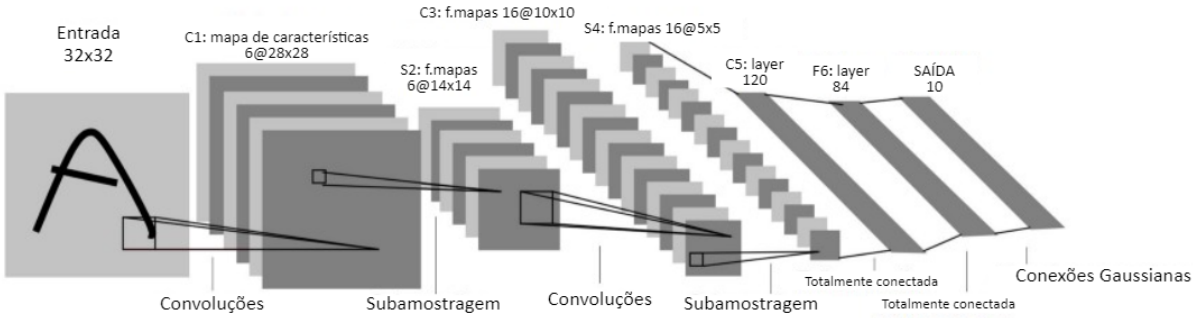
Figura 7 – Operação de *pooling*.



Fonte: (WANG et al., 2018).

Para entender melhor como todas essas camadas são conectadas ao longo de uma CNN, a Figura 8 mostra a arquitetura da CNN LeNet5 proposta por LeCun et al. (1998), composta por 2 camadas convolucionais, 2 camadas de *pooling* e 2 camadas totalmente conectadas.

Figura 8 – Arquitetura da rede LeNet5.

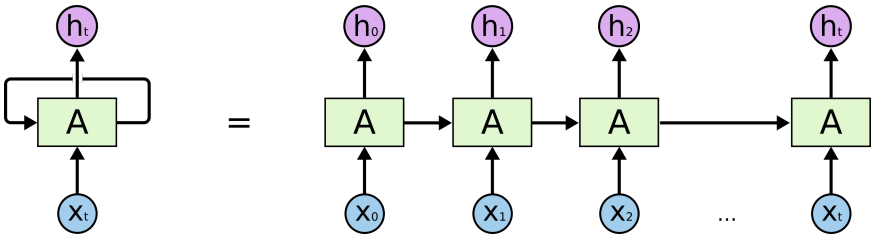


Fonte: Adaptado de (LECUN et al., 1998).

2.3 Redes Neurais Recorrentes

As redes neurais convolucionais, bem como as redes neurais profundas tradicionais não conseguem lidar com informações temporais dos dados de entrada. Para este tipo de tarefa, onde as características temporais devem ser analisadas, como texto, áudio e vídeo, as Redes Neurais Recorrentes são predominantes. O objetivo das redes neurais recorrentes é imitar como o cérebro humano usa as informações do passado para tomar decisões do presente (LIPTON; BERKOWITZ; ELKAN, 2015; DUARTE, 2019). A Figura 9 exemplifica como as saídas das camadas de uma RNN alimentam a próxima camada em seguida.

Figura 9 – Rede Neural Recorrente.



Fonte: (JUNIOR, 2019).

Segundo Bengio, Simard e Frasconi (1994), as redes neurais recorrentes possuem três requisitos básicos, que são:

- O sistema deve ser capaz de armazenar informações por um período arbitrário;
- O sistema deve ser resistente a ruídos;
- Os parâmetros do sistema devem ser treináveis.

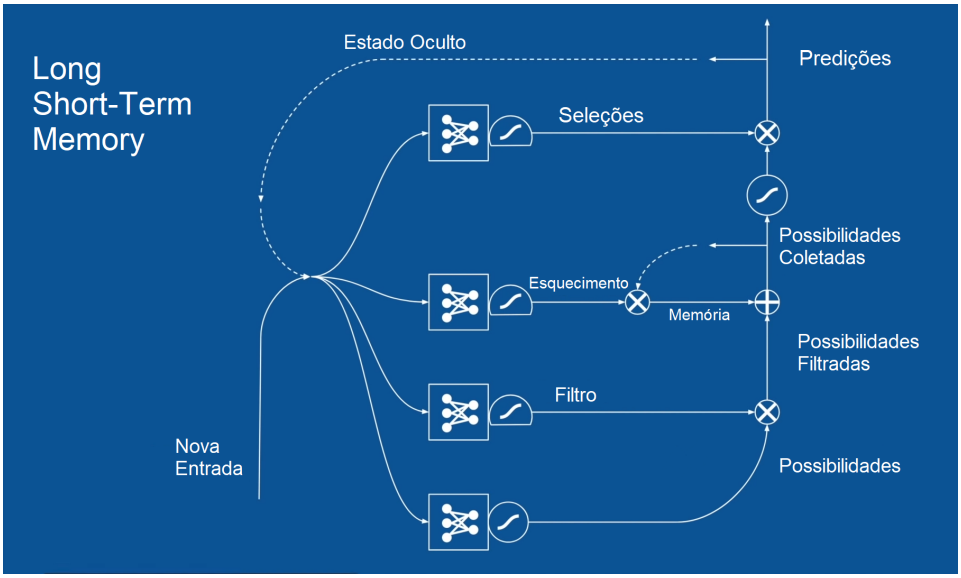
2.3.1 LSTM

Hochreiter e Schmidhuber (1997) apresentaram as redes *Long Short-Term Memory* (LSTM) que são uma extensão das redes neurais recorrentes. Elas funcionam com mecanismos de memória interna para aprender e descrever uma respresentação complexa de dependências a longo prazo entre os dados sequenciais de entrada, sendo propícias para serem usadas em problemas onde se tem características que carregam informações temporais.

Segundo Goodfellow, Bengio e Courville (2016), uma LSTM tem sua estrutura dividida em portões (*gates*) responsáveis por fazer a seleção das informações que serão escolhidas pela célula, uma vez que pode, por exemplo, ser útil para uma rede neural esquecer o antigo estado. Por exemplo no caso de ser necessário gerar um novo texto, esses portões seriam úteis para que a RNN tenha em sua 'memória' que uma palavra já foi escrita, evitando assim o acontecimento de repetições.

A Figura 10 mostra a arquitetura de uma LSTM. Como pode-se observar, uma LSTM possui quatro portões: o portão das possibilidades (*Possibilities Gate*), o portão de filtragem (*Ignoring Gate*), o portão do esquecimento (*Forgetting Gate*) e o portão das seleções (*Selection Gate*) (ROHRER, B., 2017).

Figura 10 – Arquitetura de uma LSTM.



Fonte: Adaptado de (ROHRER, B., 2017).

O portão das possibilidades é responsável por determinar quais são os possíveis *outputs* para a iteração atual. O portão de filtragem é responsável por selecionar quais entradas serão ignoradas naquela iteração, filtrando as informações irrelevantes para aquela etapa e gerando, deste modo, as possibilidades filtradas.

O portão de esquecimento é responsável por fazer a remoção de informações que não são mais úteis. Ele recebe o vetor de entradas do passo atual e o *hidden-state* do passo anterior,

multiplica a entrada pelas matrizes de peso, adiciona o *bias* e executa a função de ativação que, nesse caso, resulta em uma saída binária onde valores com saída 0 serão esquecidos e valores com saída 1 permanecem para os procedimentos seguintes.

No passo posterior, a saída fornecida pelo portão de esquecimento passa por uma operação de multiplicação elemento-por-elemento, denotada pelo sinal $\otimes$, com o vetor das possibilidades coletadas. O resultado dessa multiplicação é um vetor de valores a serem lembrados, o que pode ser chamado de memória.

Após as possibilidades serem filtradas, elas passam por uma operação de soma elemento-por-elemento com a memória, denotado pelo sinal $\oplus$, que resulta em um novo vetor de possibilidades, as possibilidades coletadas, que representam o conjunto das informações importantes para o passo atual, ou seja, as informações que foram lembradas pela memória somadas às possibilidades importantes para o passo atual e as não ignoradas pelo *Ignoring Gate*. Deste modo, as possibilidades coletadas são um conjunto de possíveis respostas mais plausível para o passo atual.

Após esse procedimento, as possibilidades coletadas passam pela função de ativação tangente hiperbólica (Tanh), responsável por ajustar os valores coletados para valores entre -1 e 1. O resultado dessa operação é então somado às possibilidades selecionadas pelo portão das seleções que recebe como entrada o vetor de entradas do passo atual e o *hidden-state* do passo anterior e calcula as probabilidades de cada possibilidade. A saída do portão das seleções passa então por um novo procedimento de multiplicação elemento-por-elemento com o vetor de *output* da função Tanh. Esse procedimento resulta nas probabilidades de cada possibilidade, que informa, deste modo, a predição, o *output*, para a célula atual. Esse *output*, também conhecido como estado oculto (*hidden-state*), é enviado para a célula seguinte e o procedimento se repete.

Uma vez que as LSTMs trabalham com dados sequenciais, pode-se usá-las para fazer a classificação de ações de acordo com as poses 2D extraídas dos quadros de um vídeo, uma vez que as ações podem ser caracterizadas por uma sequência de poses 2D.

2.4 Mecanismos de Atenção

Uma estratégia que vem sendo cada vez mais investigada e utilizada nas áreas de Visão Computacional, Aprendizado de Máquina e Reconhecimento de Padrões é a análise de atenção. Assim como as redes neurais são consideradas uma tentativa de replicar as ações do cérebro humano de uma maneira simplificada, o mecanismo de atenção também é uma tentativa de implementar a ação de concentração seletiva em algumas coisas relevantes, enquanto ignora outras.

O primeiro modelo com análise de atenção foi proposto por Bahdanau, Cho e Bengio

(2014) para melhorar o problema de tradução de textos com redes neurais, utilizando atenção para realizar uma busca automática para prever uma palavra tendo outras partes da frase origem como relevantes.

Para ilustrar como o mecanismo funciona, considere uma foto de uma turma da escola, onde tipicamente existem várias crianças sentadas em linhas e o professor sentado em algum lugar entre elas. Caso alguém pergunte quantas pessoas existem na foto, basta contar o número de cabeças e terá a resposta. Mas, caso alguém pergunte onde está o professor na foto, seu cérebro sabe exatamente o que fazer e em que características focar a atenção para identificar um adulto na foto, ignorando todo o resto. Este é o mecanismo de atenção utilizado pelo sistema visual humano para focar nos objetos de interesse em uma imagem.

Os mecanismos de atenção foram apresentados por Bahdanau, Cho e Bengio (2014) como uma forma de expandir a solução de seu trabalho anterior sobre modelos *encoder-decoder*. Esse estudo serviu de base para o trabalho de mecanismos de atenção de Vaswani et al. (2017) sobre como esta tecnologia revolucionou a área do aprendizado profundo com o conceito de processamento paralelo de palavras ao invés de processá-las sequencialmente para tarefas de tradução.

Métodos para desviar a atenção para as regiões mais importantes de uma imagem e desconsiderar as menos importantes são chamados de mecanismos de atenção. Os seres humanos fazem uso de mecanismos parecidos no dia a dia, conseguindo distinguir o que importa ou o que está procurando em determinada imagem que chega ao seu cérebro.

Os mecanismos de atenção tornaram-se parte integrante da modelagem de sequência e modelos de transdução convincentes em várias tarefas, permitindo a modelagem de dependências sem considerar sua distância nas sequências de entrada ou saída. Normalmente é utilizada uma rede neural recorrente em conjunto com estes mecanismos de atenção (VASWANI et al., 2017).

Os mecanismos de atenção podem ser categorizados de acordo com a natureza do problema a ser tratado. Para isso, existem quatro categorias fundamentais de atenção: atenção do canal, atenção espacial, atenção temporal e atenção ramificada, e duas categorias híbridas, combinando atenção de canal/espacial e espacial/temporal. Algumas combinações de mecanismos de atenção ainda não foram exploradas. Essa categorização pode ser vista na Figura 11.

Figura 11 – Categorização de Mecanismos de Atenção.

Fonte: (GUO et al., 2021).

3 Reconhecimento de Ações Humanas

Neste capítulo apresentam-se os princípios do Reconhecimento de Ações Humanas (RAH), introduzindo seus principais conceitos e diferentes categorias de ações a serem classificadas. Apresenta-se também um estudo sobre métodos existentes na literatura para reconhecimento de ações humanas, bem como as principais bases de dados utilizadas na literatura nas mais diversas pesquisas.

3.1 Ação Humana

Sistemas de reconhecimento de ações humanas têm por objetivo identificar e classificar as ações de um ou mais indivíduos em um determinado contexto. No âmbito da visão computacional, existem diversas tarefas a serem feitas a fim de alcançar este objetivo, que começam pela captura das imagens, segmentação, identificação e rastreamento dos objetos e por fim sua respectiva classificação. Uma ação pode ser considerada como uma sequência de movimentos do corpo humano que podem envolver várias partes simultaneamente, e o reconhecimento de ações em vídeos consiste em comparar os conteúdos dos frames com padrões previamente definidos e assim rotular a ação em questão (CHENG et al., 2015).

De acordo com Aggarwal e Ryoo (2011), dependendo da complexidade do reconhecimento, existem diferentes classes em que podemos agrupar ações humanas, a saber:

- **Gesto**: Sequência de movimentos associados a um significado, feitos por membros, cabeça ou face, como por exemplo esticar o braço, sorrir, dobrar a perna;
- **Ação**: Sequência de vários gestos feitos por uma pessoa, por exemplo andar, correr ou nadar;
- **Interação**: Sequência onde um humano interage com outro humano ou com um objeto, como por exemplo duas pessoas brigando ou alguém segurando uma faca;
- **Atividade de grupo**: Atividades realizadas por mais de uma pessoa e/ou objetos, como por exemplo um jogo de futebol.

Complementando esta categorização, Vrigkas, Nikou e Kakadiaris (2015) propuseram outras duas classes de atividades humanas: comportamentos e eventos. A primeira se refere a ações físicas que estão associadas a alguma emoção ou personalidade do indivíduo. Eventos representam uma atividade de caráter social e indica a intenção social da pessoa. A representação das classes definidas em Vrigkas, Nikou e Kakadiaris (2015) está ilustrada na Figura 12, onde

as ações mais acima representam uma maior complexidade, enquanto as ações de baixo são mais simples.

Figura 12 – Categorização das ações humanas.

Fonte: Vrigkas, Nikou e Kakadiaris (2015).

.

Para se realizar o reconhecimento de ações, geralmente o processo segue dois principais componentes: representação de ação e classificação de ação. O primeiro é responsável por transformar determinada ação registrada em um vídeo em um vetor de características, ou em uma série de vetores de características. O segundo componente utiliza o vetor de características para classificar a ação. Estudos recentes mostraram redes profundas, como as redes neurais convolucionais, unificando essas duas etapas fazendo com que houvesse um ganho na performance em geral (KONG; FU, 2018).

3.2 Métodos para Reconhecimento de Ações Humanas

Como já foi mencionado, o reconhecimento de ações humanas tem uma ampla possibilidade de aplicações por meio de vigilância por vídeo, análise de conteúdo e interações humanas. Por isso, vem ganhando notoriedade como um dos tópicos mais ativos nas pesquisas em visão computacional, onde muitos trabalhos vêm sendo feitos sobre o do tema.

As técnicas propostas para resolver o problema de reconhecimento de ações podem ser divididas em dois grupos: *handcrafted*, que são técnicas onde as características são extraídas de acordo com um algoritmo pré-definido baseado no conhecimento do problema, e as técnicas

baseadas em aprendizado profundo (*deep learning*), como por exemplo extrair características derivadas de uma base de dados de imagens pelo treinamento de uma rede neural convolucional (ANTIPOV et al., 2015).

Nesta seção são apresentados métodos *handcrafted* e baseados em aprendizado profundo propostos na literatura para o reconhecimento de ações humanas que são considerados correlatos ao método proposto nesta dissertação.

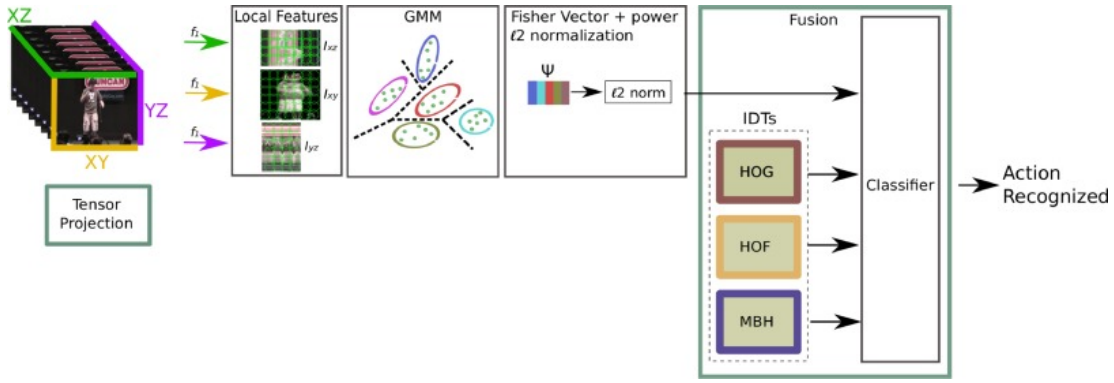
3.2.1 Métodos *Handcrafted*

Laptev (2005), a fim de detectar eventos espaço-temporais, usou a ideia dos operadores de interesse de Harris para detectar estruturas locais no tempo-espaço onde os valores das imagem tem variações locais tanto no espaço quanto no tempo. Após, extrai os descritores espaço-temporais invariantes em escala com base nas estimativas das extensões espaço-temporais dos eventos detectados.

O conceito de cubóides apresentado por Dollár et al. (2005), onde em cada ponto de interesse espaço-temporal, que são os máximos locais da resposta nos domínios espacial e temporal, um cubóide é extraído, que contém os valores de *pixel* em uma janela espaço-temporal, foi utilizado por vários métodos. Um deles, chamado Trajetórias Densas aprimoradas (iDT - *improved Dense Trajectories*), foi proposta por Wang e Schmid (2013), melhorando a versão prévia da *Dense Trajectories* (DT) (WANG et al., 2011). Baseado na construção de um cubóide usando *Bag-of-Visual-Words* (BoVW), um volume curvo criado pela anexação da vizinhança espacial em cada posição de pontos densamente amostrados que são rastreados em alguns *frames* usando fluxo óptico. Esses volumes são descritos usando sua trajetória, Histogramas de Gradientes Orientados (HoG - *Histograms of Oriented Gradients*), Histogramas de Fluxo Óptico (HoF - *Histograms of Optical Flow*) e Histogramas de Limite de Movimento (MBH - *Motion Boundary Histograms*), e então codificados usando a técnica de vetores de Fisher.

Carmona e Climent (2018) também fizeram uso de médias de Trajetórias Densas Aprimoradas, adicionando novas *features* baseadas em *templates* temporais. Esses modelos foram construídos considerando uma sequência de vídeo como um tensor de terceira ordem e computando três diferentes projeções. Podemos ver na Figura 13 o método proposto, que após a extração das *features* é utilizado uma GMM, com uma normalização, que alimenta o classificador junto aos IDTs.

Figura 13 – Método de Carmona e Climent (2018).

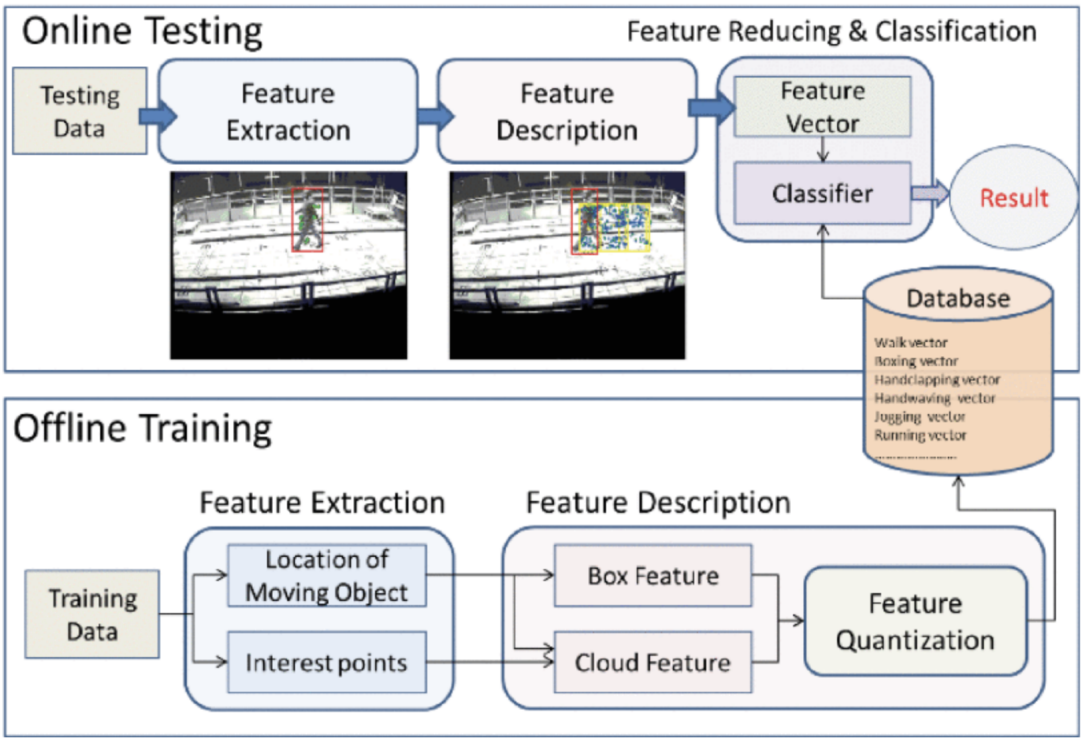


Fonte: (CARMONA; CLIMENT, 2018).

Alguns métodos fazem uso da análise da forma do corpo para o reconhecimento das ações humanas. Junejo e Aghbari (2012) apresentaram o método chamado aproximação simbólica agregada (*SAX - Symbolic Aggregate approximation*) para abordar o problema de reconhecimento de ações humanas. Utiliza-se de trajetórias de movimentos de pontos de referência de um ator como referência, que são transformados em uma representação simbólica em que não é utilizada apenas uma função de distância, mas considera juntamente parâmetros como velocidade, aceleração e curvatura. Por fim é feito o treinamento do descritor simbólico e a classificação utilizando vizinhos mais próximos (*NN - Nearest Neighbor*). O método conseguiu atingir boas acurácias para o *dataset* Weizmann, alcançando 88.6%. Apesar de não ser o estado da arte, o autor enfatiza o fato de que a estratégia é muito mais rápida e eficiente computacionalmente se comparada às outras técnicas que atingiram resultados melhores.

Chou et al. (2018) fizeram uso de silhuetas para tratar o problemas de reconhecimento de ações humanas, utilizando Modelo de Mistura Gaussiana (GMM - *Gaussian Mixture Model*) para obter as silhuetas. Em seguida são extraídos ponto de interesse, monitorando as diferenças nos *frames* baseado foco de atenção e detecção da região de interesse, onde são aplicados filtros de Gabor 2D de diferentes orientações. Após esses passos, *features* como tamanho e largura do objeto detectado e sua velocidade, bem como características espaço-temporais. Para a etapa de reconhecimento, foram avaliados três diferentes classificadores: vizinho mais próximo, classificador de Modelo de Mistura Gaussiana e média mais próxima. O sistema pode ser observado na Figura 14.

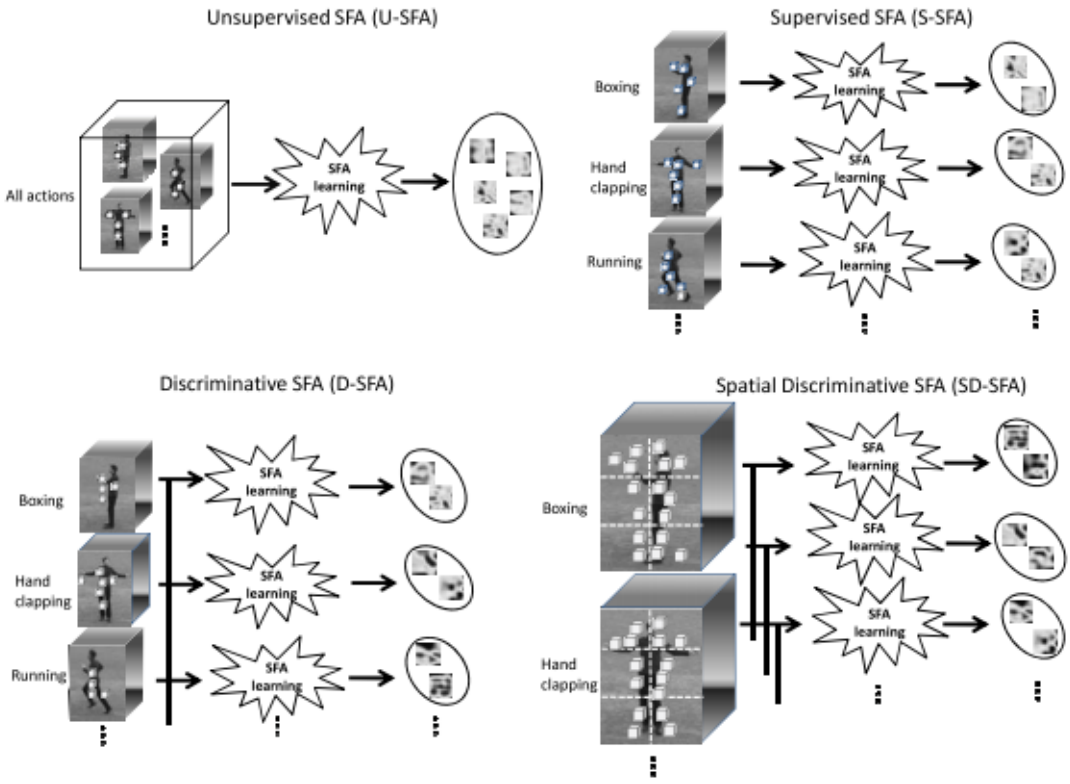
Figura 14 – Método de Chou et al. (2018).



Fonte: (CHOU et al., 2018).

Zhang e Tao (2012) propuseram um grupo de features para reconhecer ações humanas, inspiradas pelo princípio de lentidão temporal, que foi aplicado com sucesso para modelar o campo visual receptivo dos neurônios corticais. O princípio de lentidão temporal se refere ao sinal sensorial primário, que são, por exemplo, as respostas dos receptores da retina ou os valores do *pixel* cinza de uma câmera CCD (*Charge-Coupled Device*) que variam rapidamente em um curto período de tempo. No trabalho de Zhang e Tao (2012) existem quatro passos principais para o reconhecimento de ações humanas baseado em análise lenta de features: primeiro, cubóides de momentos aleatórios do movimento são extraídos, onde cada cubóide é obtido de uma sequência de três quadros; posteriormente são usadas diferentes estratégias de aprendizado para extrair as características de forma lenta, mostrada na Figura 15; o terceiro passo é transformar as features obtidas no passo anterior para gerar *features* ASD (*Accumulated Squared Derivative*), que codifica a distribuição estatística de características lentas em uma sequência de ação; e por fim, é treinada uma SVM (*Support Vector Machine*) para classificar as ações humanas representadas pelas *features* ASD. A estratégia de aprendizado que obteve o melhor resultado para as base de dados KTH e Weizmann foi a *Spatial Discriminative SFA* (SD-SFA), obtendo 93.5% e 93.87% de acurácia, respectivamente.

Figura 15 – Estratégias de aprendizado para o SFA.



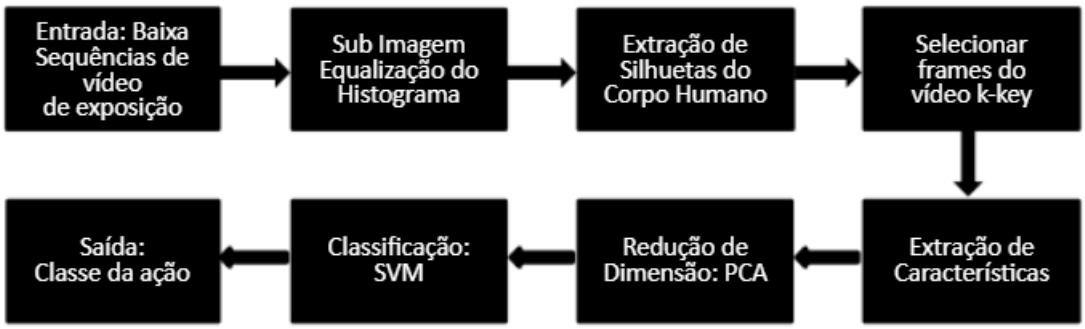
Fonte: (ZHANG; TAO, 2012).

Guo, Ishwar e Konrad (2013) propuseram a utilização de matrizes de covariância empíricas de *features*. Uma matriz de covariância empírica para representar a ação de forma compacta é formada a partir da descrição localizada da ação, a qual é calculada por um conjunto de vetores de características espaço-temporais extraídos a partir do vídeo. Usando esta matriz, foram desenvolvidos dois métodos de aprendizado para reconhecimento de ações: o primeiro utiliza a classificação do vizinho mais próximo usando uma métrica Riemanniana adequada para matrizes de covariância e o segundo método usa a aproximação linear esparsa fazendo o uso do logaritmo da matriz de covariância. O método alcançou incríveis 100% e 98.5% nos *datasets* Weizmann e KTH, respectivamente.

Chaaraoui, Climent-Pérez e Flórez-Revuelta (2013) utilizaram um método baseado nos pontos de contorno da silhueta humana. Ao extrair os pontos da silhueta, também é calculada o centro de massa da mesma para se calcular a distância normalizada entre as duas informações, a fim de fazer com que as variações de escala de silhuetas não sejam um problema. Outro aspecto importante do trabalho é o fato de serem utilizadas poses-chaves para respresentar as ações. Também é utilizado DTW (*Dynamic Time Warping*) para tornar o problema invariante ao começo e fim da ação, uma vez que cada pessoa pode executar um movimento em diferentes velocidades. Por fim, os dados das silhuetas para cada ação são concatenados, formando o descritor final, e a classificação é dada por um simples classificador de vizinho mais próximo.

Singh e Vishwakarma (2019) também usaram um método baseado em extração de silhuetas, que consistia na seleção de *k-poses* chaves de uma saída melhorada derivada de sequências de entradas de vídeo de baixa qualidade usando uma técnica de sub-histograma de refinamento de baixa exposição. A silhueta do corpo humano é extraída dessas estruturas de equalização usando a técnica de segmentação GLCM (*Gray-Level Co-occurrence Matrix*) e normalizado em células fixas. Por fim, é feita uma redução de dimensionalidade utilizando PCA (*Principal Component Analysis*) e a classificação da ação é feita por meio do classificador linear SVM, como mostra a Figura 16.

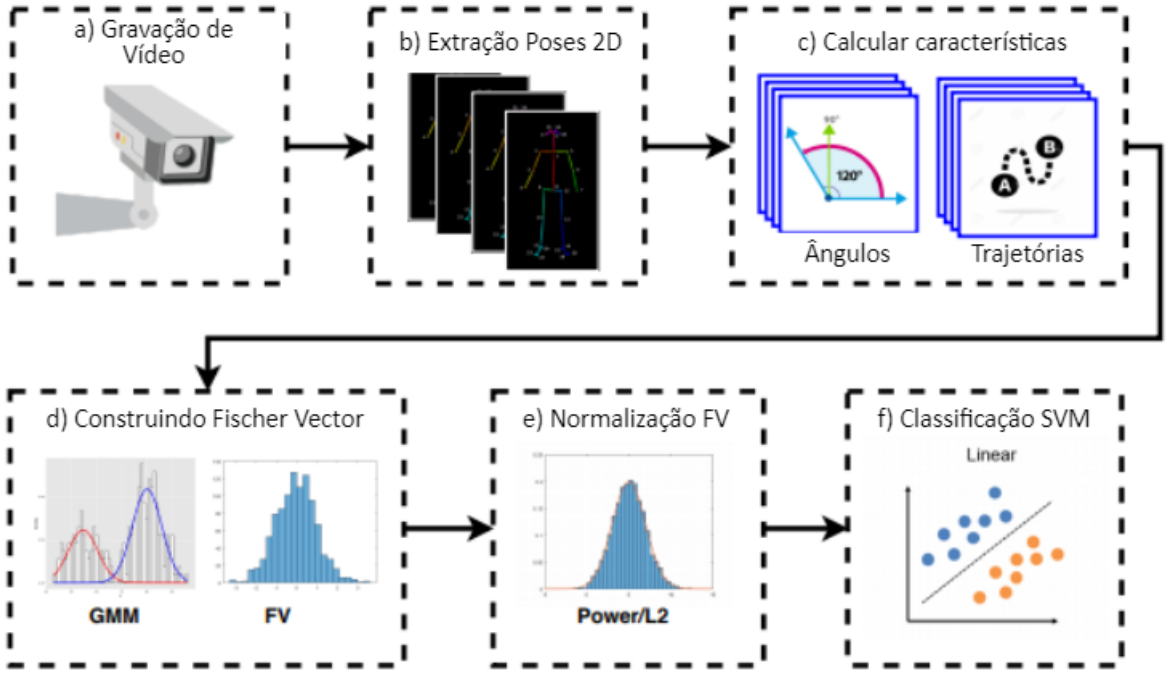
Figura 16 – Método de Singh e Vishwakarma (2019).



Fonte: Adaptado de (SINGH; VISHWAKARMA, 2019).

Silva (2020) fez o uso do OpenPose para gerar poses 2D e extrair *features* que descrevam aquela pose com base nos ângulos formados entre os segmentos de linha dos esqueletos e também com base na trajetória desses pontos ao longo dos frames do vídeo. Com as *features* extraídas, foi utilizado um *Fisher Vector* para codificar-lás e posteriormente foi aplicada uma normalização ao vetor. Finalmente, uma SVM linear foi utilizada para a classificação. O método proposto pode ser visto na Figura 17.

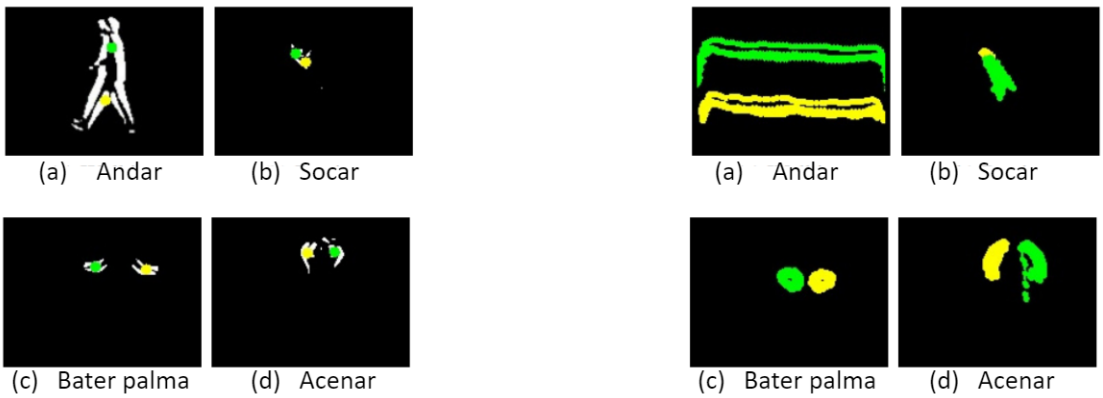
Figura 17 – Método de Silva (2020).



Fonte: Adaptado de (SILVA, 2020).

Doumanoglou, Vretos e Daras (2016) aplicaram um filtro Sobel 3D no vídeo resultando em uma imagem binária com *pixels* de valores diferentes de zero na área de movimento. Os valores desses *pixels* são espacialmente agrupados utilizando *k-means* e os centros de movimento mais dominantes do vídeo são extraídos. Os centros então são rastreados formando trajetórias esparsas que são utilizadas para criar um novo tipo de *feature* chamado Histograma de Trajetórias Orientadas (HOT - *Histogram of Oriented Trajectories*), descrevendo o vídeo. Os vetores de características são então passados para um classificador AdaBoost para realizar a classificação, como é representado na Figura 18.

Figura 18 – Método de Doumanoglou, Vretos e Daras (2016).



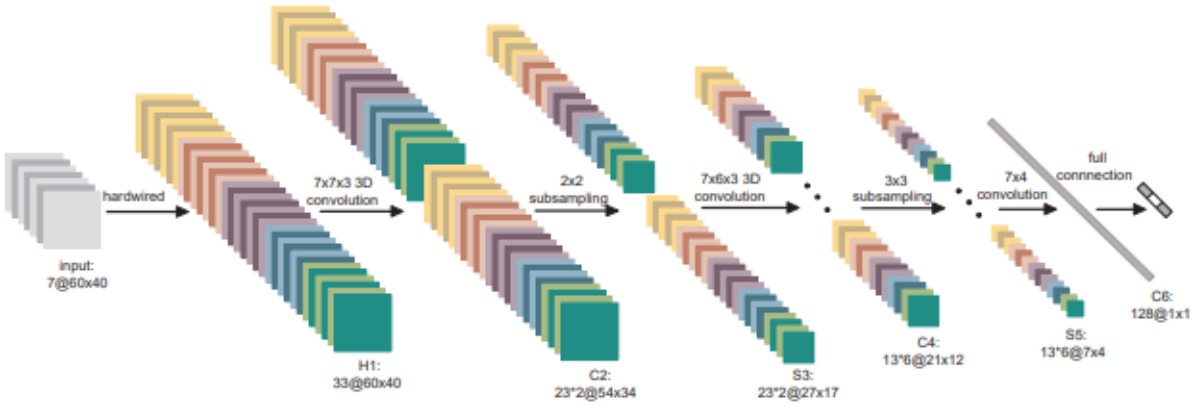
Fonte: Adaptado de (DOUMANOGLOU; VRETOS; DARAS, 2016).

Moreira, Menotti e Pedrini (2020) apresentaram uma metodologia de descrição de volume utilizando a representação Visual Rhytim (VR), que consiste em codificar os vídeos em imagens, juntando fatias de cada *frame* do vídeo, que são horizontalmente concatenadas formando uma imagem de dimensões *W x H*, onde *W* é a duração em *frames* e *H* é o tamanho em *pixels* da fatia. Dessa forma, cada coluna representa um instante no tempo e cada linha representa um *pixel* da imagem. A partir disso, dois métodos de extração de caracteríticas são propostos, onde a primeira é a aplicação direta da técnica no vídeo original, e a segunda foca na análise de pequenas vizinhanças obtidas a partir do processo de trajetórias densas.

3.2.2 Métodos Baseados em Redes Neurais Convolucionais

Ji et al. (2012) fizeram uso de Redes Neurais Convolucionais para poder agir diretamente em entradas brutas, além de automatizar o processo de construção de *features*. Com a utilização de uma 3D CNN, são extraídas características de dimensões espaciais e temporais, realizando convoluções 3D, assim capturando informações de movimento codificadas em vários *frames* adjacentes do vídeo. O modelo gera múltiplos canais de informação dos *frames* de entrada, e a representação final da *feature* é dada pela combinação dos canais. A arquitetura utilizada pode ser visualizada na Figura 19.

Figura 19 – Arquitetura da CNN de Ji et al. (2012).

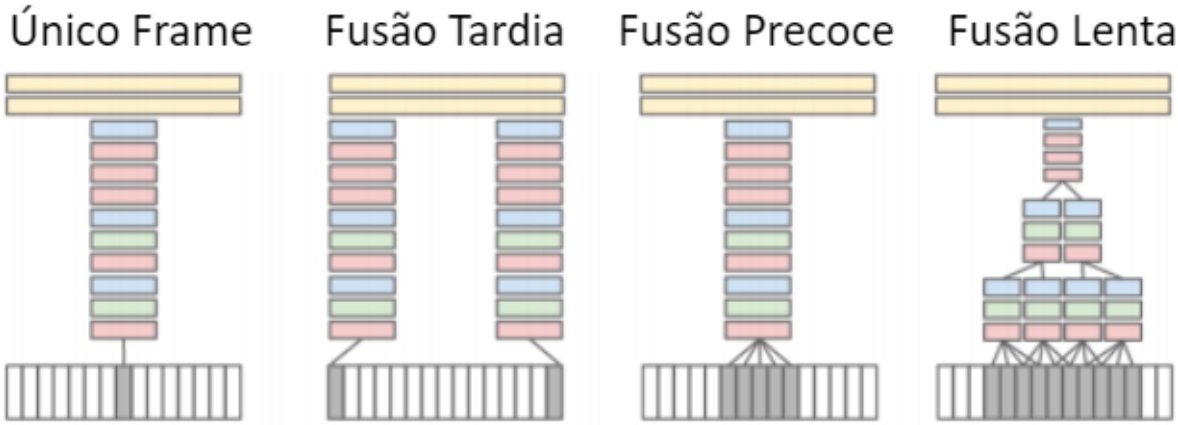


Fonte: (JI et al., 2012).

Karpathy et al. (2014) fizeram uso de CNN para extrair informações temporais com quatro diferentes estratégias, representadas na Figura 20: a primeira (*Single Frame*) consiste em utilizar uma arquitetura de *frames* únicos para entender a contribuição da aparência estática para a precisão da classificação; a segunda (*Early Fusion*) modifica os filtros da primeira camada convolucional da estratégia anterior estendendo-os para serem de tamanho $11 \times 11 \times 3 \times T$ *pixels*, onde *T* é uma dimensão temporal; a terceira (*Late Fusion*) faz uso de duas redes *Single Frame* que são combinadas na primeira camada totalmente conectada; já a quarta (*Slow*

Fusion) mistura as duas abordagens, fazendo com que as camadas superiores tenham acesso a informações mais globais em ambas dimensões temporais e espaciais.

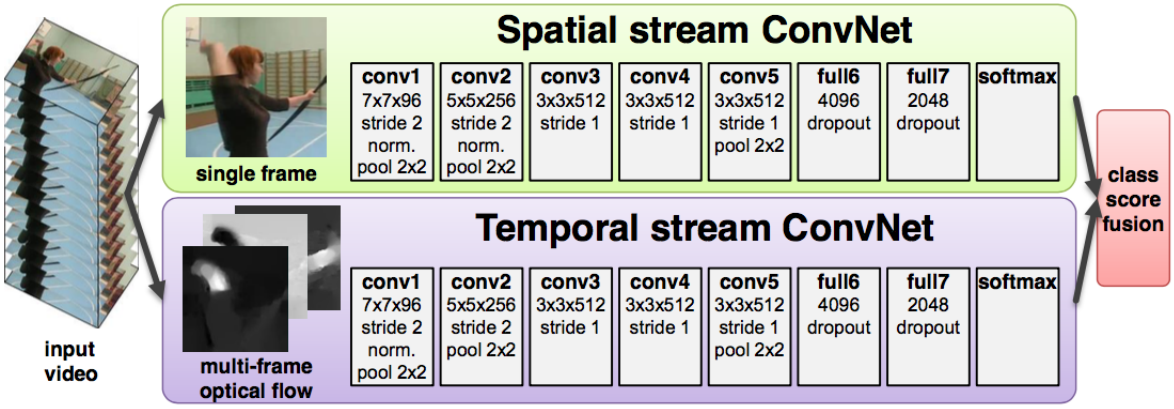
Figura 20 – Estratégias de CNN de Karpathy et al. (2014).



Fonte: Adaptado de (KARPATHY et al., 2014).

A ideia de usar duas redes para reconhecimento de ações humanas foi introduzida por Simonyan e Zisserman (2014), onde o objetivo principal era incorporar uma rede espacial, que extrai informações da aparência dos *frames* e carrega informações sobre as cenas e objetos retratados no vídeo, junto a uma rede temporal que extrai a forma do movimento ao longo dos *frames* e repassa a informação do movimento da câmera e dos objetos, como pode ser visto na Figura 21.

Figura 21 – Arquitetura de Simonyan e Zisserman (2014).



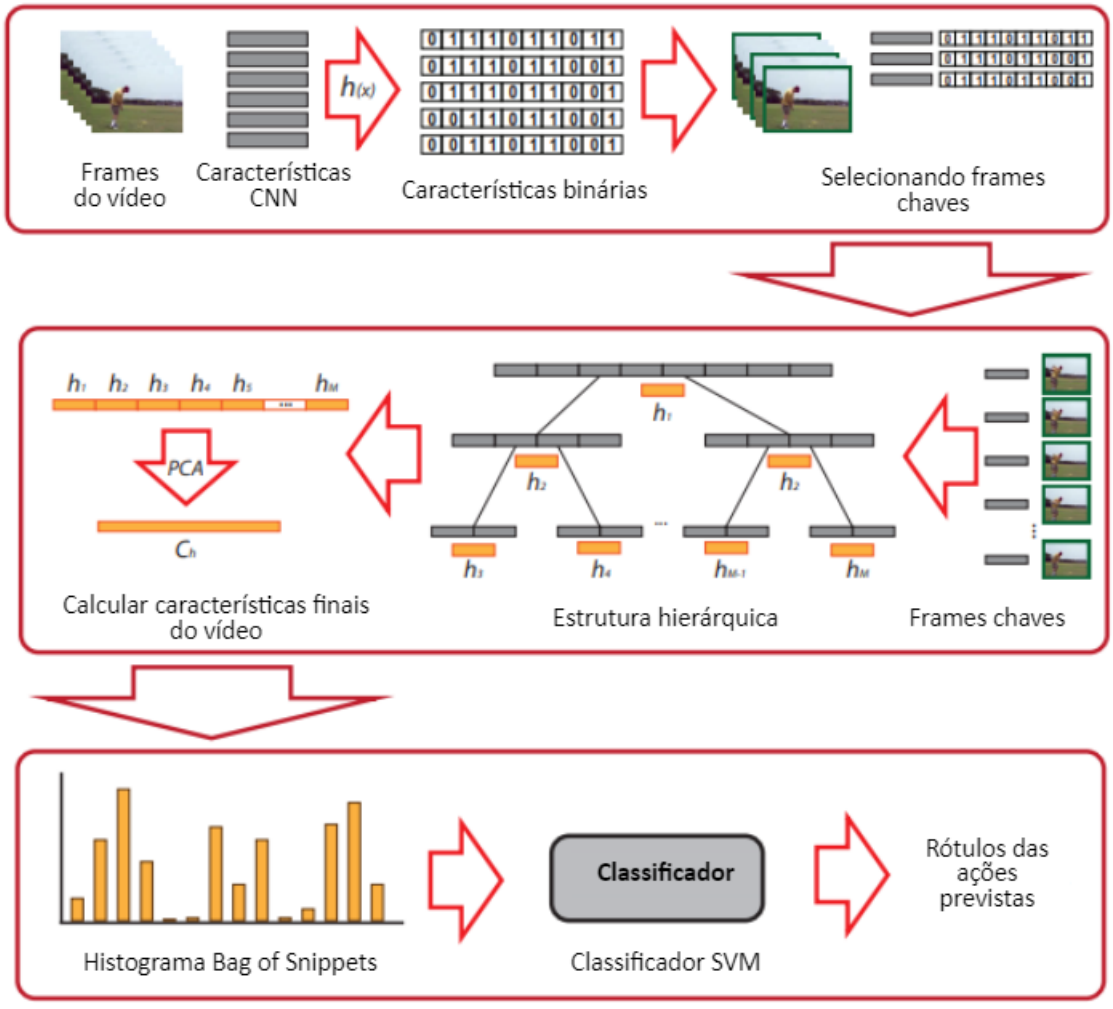
Fonte: (SIMONYAN; ZISSERMAN, 2014).

Baseado na arquitetura proposta por Simonyan e Zisserman (2014), que usa redes espaciais e temporais, Carreira e Zisserman (2017) utilizaram filtros e *pooling kernels* para

expandir a arquitetura de 2D CNN *inflation* e criar uma nova arquitetura de duas redes chamada Inflated 3D CNN (I3D).

Ravanbakhsh et al. (2015) também fizeram uso de redes neurais convolucionais para extrair as *features* utilizadas para classificar a ação. O método consiste em, dado um vídeo, extrair as *features* derivadas da última camada de uma rede neural convolucional para cada *frame* e então mapeá-las em um espaço de código binário, então rastrear mudanças em cada *bit* do código ao longo do tempo para selecionar frames-chaves. Trechos de vídeos entre frames-chaves alimentam uma decomposição hierárquica, onde é aplicado um PCA em cada nível dessa hierarquia para reduzir a dimensão. Todos os níveis são empilhados como uma representação vetorial para um trecho do vídeo e por fim é construído um histograma de palavras temporais para todos os vídeos para treinar um classificador para fazer a predição das ações. Todo o fluxo desde o *input* até a predição das ações do método é representado na Figura 22.

Figura 22 – Método de Ravanbakhsh et al. (2015).

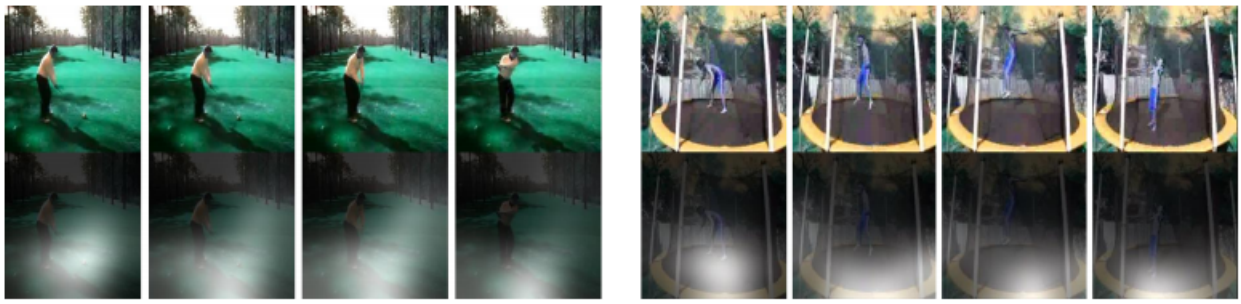


Fonte: Adaptado de (RAVANBAKHSH et al., 2015).

3.2.3 Métodos Baseados em Análise de Atenção e Redes Neurais Recorrentes

Sharma, Kiros e Salakhutdinov (2015) propuseram um modelo baseado em atenção para a tarefa de reconhecimento de ações em vídeos, fazendo o uso de Redes Neurais Recorrentes com LSTM, onde o modelo aprende quais partes do *frame* são mais relevantes para a tarefa em questão e atribui maior importância a elas. Na Figura 23 por exemplo, as partes mais claras foram as regiões da imagem que o modelo entendeu que devesse ter mais atenção na hora de realizar o treinamento.

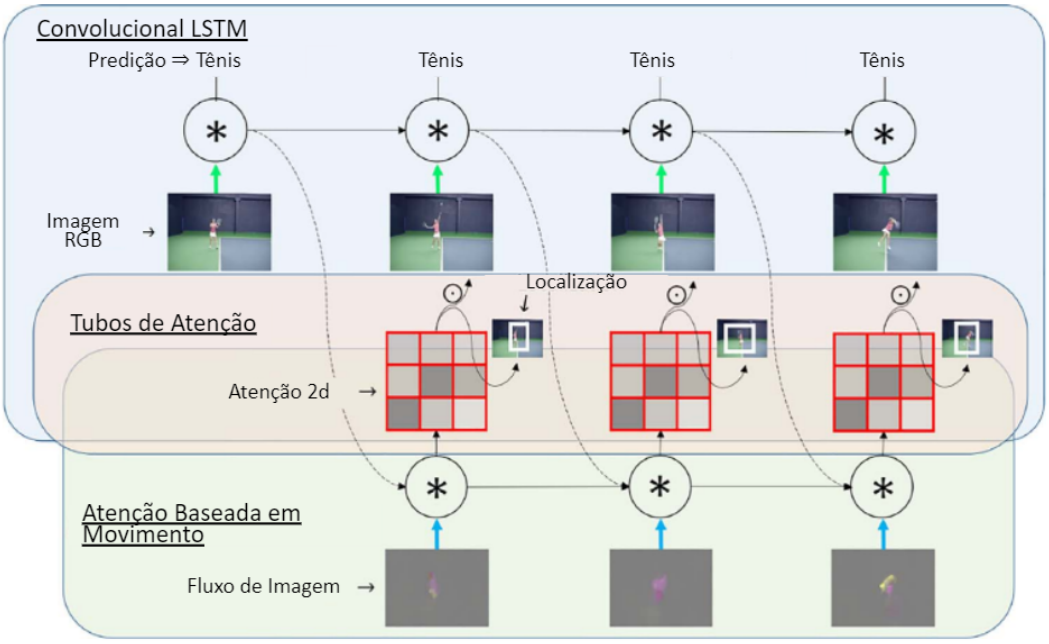
Figura 23 – Análise de atenção de Sharma, Kiros e Salakhutdinov (2015).



Fonte: (SHARMA; KIROS; SALAKHUTDINOV, 2015).

Na mesma linha, Li et al. (2018) usaram uma arquitetura que faz uso de uma *soft-attention* LSTM com a adição de convoluções e trabalhando com a atenção baseada em movimentos, onde não apenas informa sobre o conteúdo da ação, mas orienta para as localizações relevantes. A arquitetura da rede proposta pode ser vista na Figura 24.

Figura 24 – Arquitetura de Li et al. (2018).

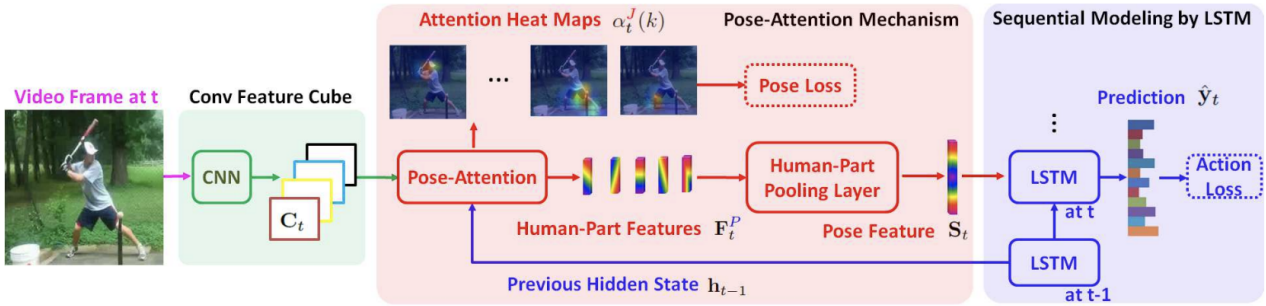


Fonte: Adaptado de Li et al. (2018).

.

Du, Wang e Qiao (2017) propuseram uma *Recurrent Pose-Attention Network* (RPAN) onde foi introduzido um novo mecanismo de análise de atenção para pose para aprender de forma adaptativa características relacionadas à pose em cada previsão de ação em cada intervalo de tempo da RNN. Em detalhes, a rede proposta é uma rede recorrente de ponta a ponta que pode explorar importantes evoluções espaço-temporais da pose para auxiliar no reconhecimento da ação humana. Esta rede aprende características compartilhando parâmetros de atenção parcialmente nas articulações humanas semanticamente relacionadas, em vez de aprender características individuais das articulações separadas. O método pode ser visto de ponta a ponta na Figura 25.

Figura 25 – Método de Du, Wang e Qiao (2017)



Fonte: (DU; WANG; QIAO, 2017).

.

3.3 Bases de Dados de Ações Humanas

Na literatura, há várias bases de dados de ações disponíveis para o reconhecimento de ações humanas. Em algumas delas, a forma de se referir às ações variam desde as atividades até os movimentos. A Tabela 1 compara algumas características das bases de dados que serão apresentadas a seguir, que são: número de classes possíveis, número de câmeras utilizadas, número de indivíduos na mesma imagem, a vista em que os vídeos foram gravados e o fundo utilizado.

Tabela 1 – Características das bases de dados disponíveis para o reconhecimento de ações humanas.

Base de dados	Nº de classes	Nº de câmeras	Nº de indivíduos	Vista	Fundo	Ano	Nº de vídeos
KTH	6	1	1	Horizontal	Homogêneo	2004	599
Weizmann	10	1	1	Horizontal	Estático	2007	93
Volleyball	9 individuais/8 coletivas	1	12	Horizontal	Dinâmico	2016	55
MuHAVi	17	8	1	Diversas	Estático	2010	119
IXMAS	13	5	1	Diversas	Estático	2006	429
URADL	10	1	1	Horizontal	Estático	2009	150
MSR Action	3	1	1	Horizontal	Dinâmico	2011	70
UCF-Sports	9	1	1	Variadas	Dinâmico	2014	182
UT-Interaction	6	1	2	Horizontal	Estático	2009	20
LIRIS	10	1	Variável	Horizontal	Estático	2014	167
UTKinect-Action3D	10	1	1	Horizontal	Estático	2012	200

3.3.1 KTH

A base de dados KTH (SCHULDT; LAPTEV; CAPUTO, 2004) contém seis classes de ações humanas, que são: andar, trotar, correr, socar, acenar com as mãos e bater palmas. Cada uma dessas ações foi realizada por 25 pessoas em quatro diferentes cenários, ao ar-livre, ao ar-livre com variação de escala, ao ar-livre com roupas diferentes e em locais fechados, como é mostrado na Figura 26. A base de dados contém 599 vídeos adquiridos em fundos similares com uma câmera estática, somando um total de 289.715 frames, 11.375,32 segundos, capturados em 25 *frames* por segundo (FPS) e de tamanho 160×120 *pixels*.

Figura 26 – Images de ações humanas obtidas de vídeos da base de dados KTH.

Fonte: Adaptado de Schuldt, Laptev e Caputo (2004).

3.3.2 Weizmann

A base de dados Weizmann (GORELICK et al., 2007a) consiste em 10 classes, que são andar, correr, saltar lateralmente, abaixar, acenar com uma mão, acenar com duas mãos, pular no lugar, pular estrela, pular com os dois pés e pular com um pé só, que são realizadas por 9 diferentes atores, algumas vezes mais de uma vez, resultando em 93 vídeos. A base de dados possui um total de 5.701 *frames*, 228.04 segundos, capturados em 25 FPS e tamanho 180×144 *pixels*. Todos as ações ocorrem no mesmo fundo estático como é mostrado na Figura 27.

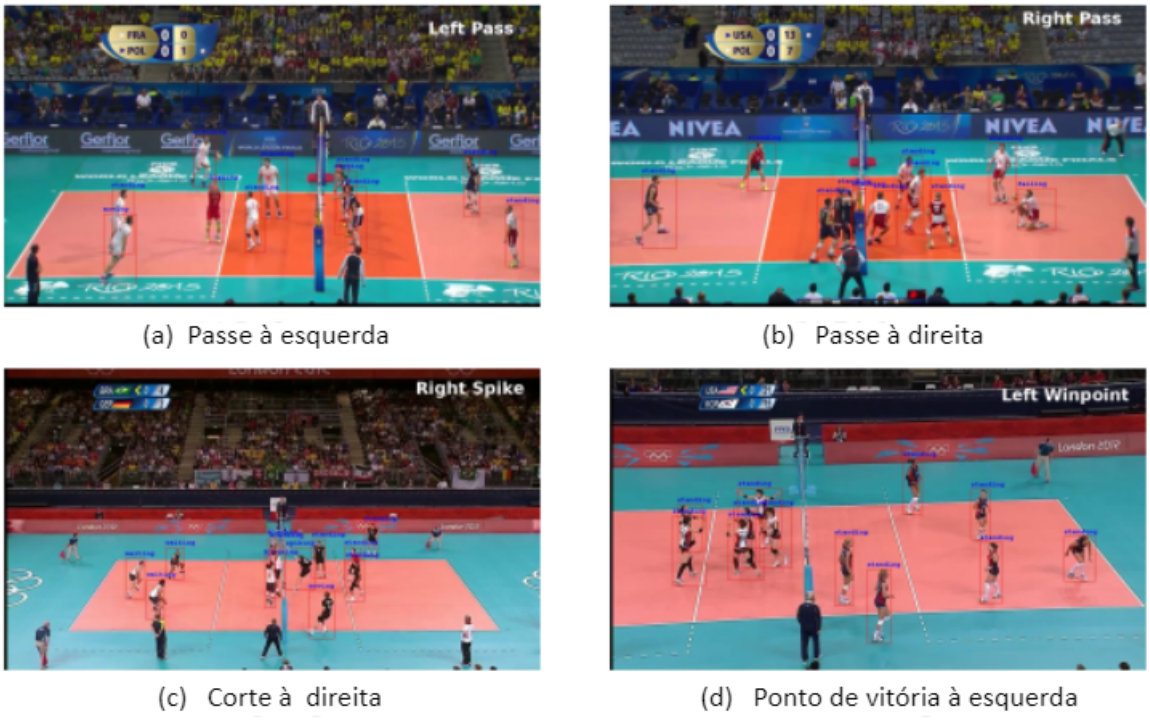
Figura 27 – Imagens de ações humanas obtidas de vídeos da base de dados Weizmann.

Fonte: Adaptado de Gorelick et al. (2007a).

3.3.3 Volleyball

A base de dados Volleyball é uma das poucas disponíveis publicamente para reconhecimento de ações multi-pessoas que é relativamente grande e contém rótulos com as posições das pessoas e também suas ações individuais e coletivas. Essa base de dados consiste em 55 vídeos de jogos de vôlei, com resolução 1920×1080 ou 1280×720 *pixels*. Ao total são 4830 *frames* rotulados, onde cada jogador é anotado com uma *bounding box* e rotulado com uma das 9 ações individuais: Esperando, Sacando, Defendendo, Caindo, Cortando, Bloqueando, Pulando, Movendo ou Parado e as cenas são rotuladas com uma das 8 ações coletivas: Levantamento a direita, Corte a direita, Passe a direita, Ponto a direita, Levantamento a esquerda, Corte a esquerda, Passa a esquerda e Ponto a esquerda, que define qual parte do jogo está acontecendo. Para cada *frame* anotado, existem múltiplos *frames* ao redor não anotados disponíveis (20 *frames* antes e 20 depois). A Figura 28 mostra quatro *frames* selecionados dessa base de dados ilustrando quatro ações coletivas durante a partida, em que cada jogador tem sua própria *bounding box* e seu respectivo rótulo para sua ação individual.

Figura 28 – Imagens de ações humanas obtidas de vídeos da base de dados Volleyball.



(a) Passe à esquerda (b) Passe à direita

(c) Corte à direita (d) Ponto de vitória à esquerda

Fonte: Adaptado de Ibrahim et al. (2016).

3.3.4 MuHAVi

A base MuHAVi (SINGH; VELASTIN; RAGHEB, 2010) é constituída de 17 classes de ações: subir escadas, ficar de joelhos, pular um obstáculo, chutar, grafitar, olhar em um carro, rastejar de joelhos, andar bêbado, pegar e jogar objeto, puxar objetos pesados, dar um soco,

reagir ao impacto de tiro de arma de fogo, parar de correr, destruir um objeto, caminhar e cair, caminhar e voltar a acenar com os braços. Cada uma das ações foram realizadas por 7 atores em 119 vídeos em um ambiente fechado com fundo não homogêneo. As capturas foram feitas por 8 câmeras dispostas em diferentes posições, como mostrado nos exemplos da Figura 29.

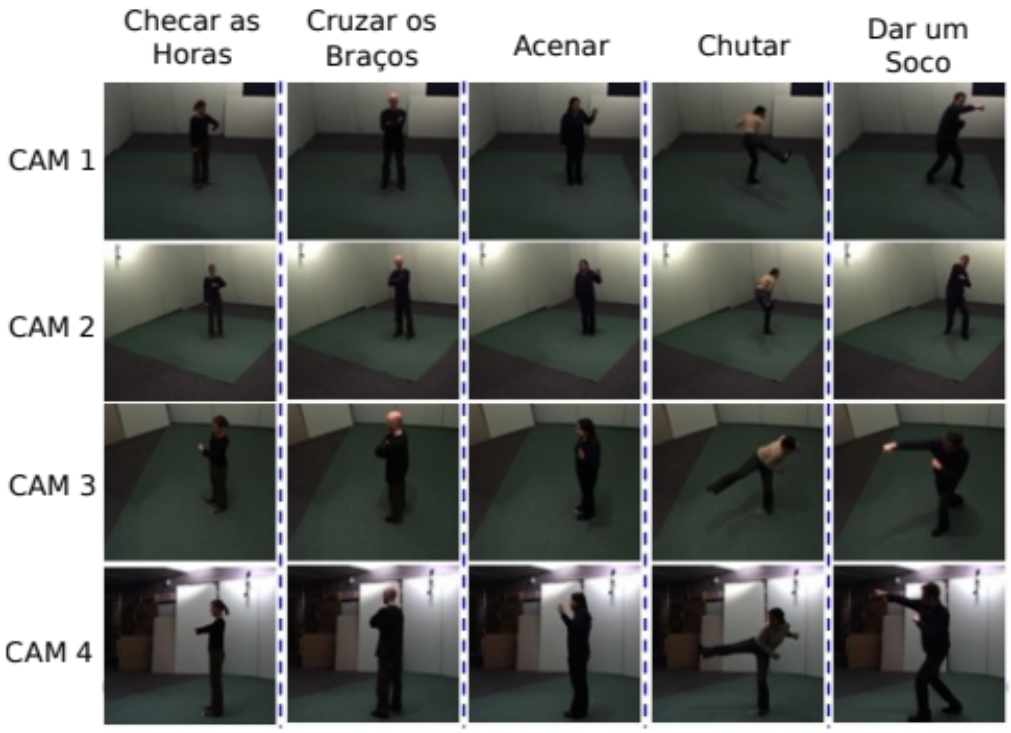
Figura 29 – Imagens de ações humanas obtidas de vídeos da base de dados MuHAVi.

Fonte: (SINGH; VELASTIN; RAGHEB, 2010).

3.3.5 IXMAS

A base IXMAS (WEINLAND; RONFARD; BOYER, 2006) contém 13 classes de ações: verificar as horas, cruzar os braços, coçar a cabeça (em um sinal de dúvida), sentar, levantar, dar a volta, andar, acenar, dar um soco, chutar, apontar, pegar algo e arremessar algo de cima para baixo ou de baixo para cima. Ainda há a existência do estado "nenhuma ação", que sugera que o ator está parado ou realizando alguma ação a qual não possui rótulo. As imagens possuem resolução de 390×291 *pixels* e a base oferece silhuetas anotadas manualmente para as pessoas na cena. A orientação dos atores podem variar de uma sequência para a outra de uma mesma ação. As capturas das imagens são feitas por 5 câmeras simultaneamente: 4 posições ortogonais e uma quinta visão superior, como pode ser visto na Figura 30.

Figura 30 – Imagens de ações humanas obtidas de vídeos da base de dados IXMAS.



Fonte: (WEINLAND; RONFARD; BOYER, 2006).

3.3.6 URADL

A base URADL (University of Rochester Activities of Daily Living) (MESSING; PAL; KAUTZ, 2009) possui 10 classes de ações: atender telefone, discar no telefone, beber água, descascar banana, comer banana, cortar banana, comer lanche, olhar a agenda telefônica, escrever no quadro branco e utilizar talheres.

A base dispõe de tomadas de ações previamente segmentadas e vídeos de alta resolução (1280×720 *pixels*) que são gravados a 30 FPS. A base dispõe ainda de vídeos curtos contendo apenas o plano de fundo para aprendizado prévio. Um exemplo de cada classe de ação pode ser visto na Figura 31. Diferentemente das outras bases de dados, as imagens capturas na URADL não dispõem de toda a silhueta do ator, uma vez que a parte inferior do corpo é coberta por uma bancada. Outra particularidade desta base é o fato de diversos objetos serem utilizados nas ações, inclusive havendo interação com o ator.

Figura 31 – Imagens de ações humanas obtidas de vídeos da base de dados URADL.

Fonte: (MESSING; PAL; KAUTZ, 2009).

3.3.7 MSR Action

A base MSR Action (YU; YUAN; LIU, 2011) da Microsoft é mais simples que as anteriores e possui apenas 3 classes de ações: bater palmas, acenar e dar socos. Os vídeos possuem resolução de 320×240 *pixels*, em dois conjuntos diferentes: um contendo 16 e outro contendo 54 vídeos. Filmados em câmera estática, estes vídeos são usados apenas para testes, já que o treinamento é realizado com os vídeos da base KTH. Alguns exemplos de quadros desta base podem ser vistos na Figura 32.

Figura 32 – Imagens de ações humanas obtidas de vídeos da base de dados MSR Action.

Fonte: (YU; YUAN; LIU, 2011).

3.3.8 UCF-Sports

A base UCF-Sports (SOOMRO; ZAMIR, 2014) possui 9 classes de ações: mergulhar, jogar golfe, chutar, levantar peso, andar a cavalo, correr, andar de skate, efetuar um lançamento de baseball e andar. As ações foram coletadas de vários canais de televisão esportivos e totalizam 182 vídeos em resolução de 720×480 *pixels* capturados em câmera estática. Na Figura 33 é possível visualizar alguns exemplos da base de dados UCF-Sports.

Figura 33 – Imagens de ações humanas obtidas de vídeos da base de dados UCF-Sports.

Fonte: (SOOMRO; ZAMIR, 2014).

3.3.9 UT-Interaction

A base UT-Interaction (RYOO; AGGARWAL, 2009) possui 6 classes de ações: cumprimentar, abraçar, chutar, dar soco, apontar e empurrar. Como se pode perceber, a base é

destinada a ações envolvendo a interação entre os indivíduos. Ao todo são 20 vídeos em câmera estática, com resolução de 720×480 *pixels*. Os vídeos estão divididos em dois conjuntos com diferentes planos de fundo. Alguns exemplos de imagens desta base podem ser vistos na Figura 34.

Figura 34 – Imagens de ações humanas obtidas de vídeos da base de dados UT-Interaction.

Fonte: (RYOO; AGGARWAL, 2009).

3.3.10 LIRIS

A base LIRIS (Laboratoire d'InfoRmatique en Image et Systèmes d'information) (WOLF et al., 2014) possui 10 classes de ações mais complexas em relação as atividades mais comuns vistas em outras bases de dados: discussão entre duas ou mais pessoas, dar um objeto para outra pessoa, pegar ou colocar um objeto em um local, entrar ou sair de um escritório, tentar entrar em um escritório e não conseguir, destrancar e entrar no escritório, deixar uma bagagem e sair, cumprimentar com aperto de mão, digitar em um teclado e falar ao telefone. Alguns exemplos de imagens desta base podem ser vistos na Figura 35.

Figura 35 – Imagens de ações humanas obtidas de vídeos da base de dados LIRIS.

Fonte: (WOLF et al., 2014).

3.3.11 UTKinect-Action3D

A base UTKinect-Action3D (XIA; CHEN; AGGARWAL, 2012) fez uso de um único Kinect estacionário para capturar os vídeos, e possui 10 ações: andar, sentar, levantar, pegar, carregar, jogar, empurrar, puxar, acenar com as mãos, bater palmas. Além das imagens em RGB, é disponibilizado imagens de profundidade em arquivo .xml e as informações das juntas dos esqueletos em um arquivo .txt. Alguns exemplos de imagens desta base podem ser vistos na Figura 36.

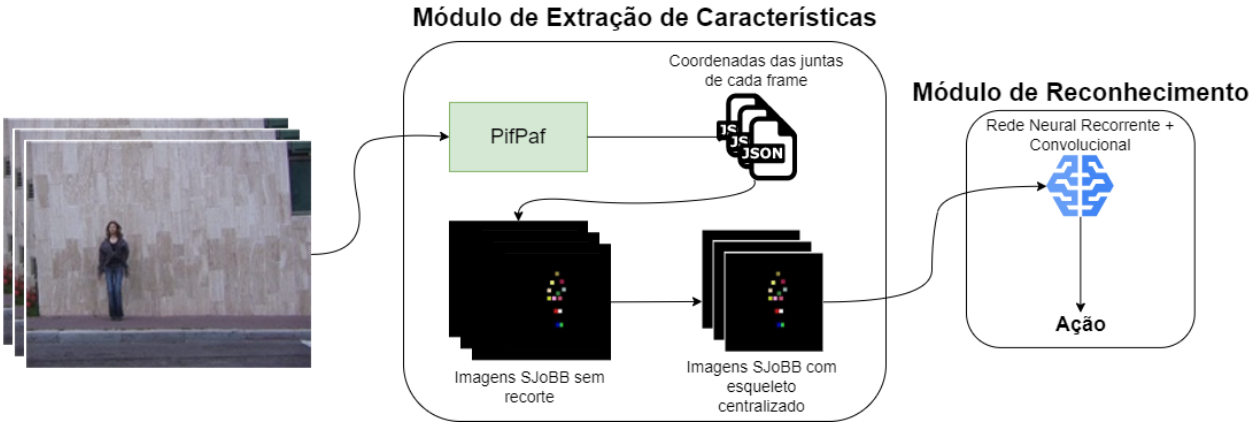
Figura 36 – Imagens de ações humanas obtidas de vídeos da base de dados UTKinect-Action3D.

Fonte: (XIA; CHEN; AGGARWAL, 2012).

4 Método Proposto

O método proposto neste trabalho para o reconhecimento de ações humanas é composto por um módulo de extração de características, que a partir da sequência de quadros do vídeo de entrada, calcula as poses 2D e gera imagens das juntas dos esqueletos em um fundo preto. Estas imagens, chamadas de *Skeleton Joints on Black Background* (SJoBB), alimentam um segundo módulo responsável por realizar a classificação da ação de acordo com as características e padões presentes nas imagens SJoBB. A Figura 37 mostra o diagrama do método proposto.

Figura 37 – Diagrama do método proposto para reconhecimento de ações humanas baseado em imagens SJoBB.



Fonte: Elaborada pelo autor.

4.1 Módulo de Extração de Características

Na etapa de extração de características, o vídeo de entrada é submetido ao processo de estimação de pose 2D pelo método PifPaf (KREISS; BERTONI; ALAHI, 2019), que gera os seguintes dados: o esqueleto (pose) da pessoa, as coordenadas das juntas do esqueleto e um mapa de calor das juntas do esqueleto. Esses dados são utilizados para gerar as imagens SJoBB, que são utilizadas no passo seguinte como características pelo módulo de reconhecimento de ações humanas.

Para gerar as imagens SJoBB é utilizado o arquivo JSON gerado pelo método PifPaf contendo as coordenadas das juntas do esqueleto. As imagens SJoBB consistem em um fundo preto com os *pixels* correspondentes às juntas coloridos com as cores próprias definidas para cada junta. Na sequência, é feito um recorte das imagens geradas de tal modo que o esqueleto fique centralizado. Finalmente, as imagens geradas são escalonadas de tal modo que todas

tenham a mesma dimensão. A Figura 38 mostra um exemplo de imagem SJoBB, de dimensão 120×120 *pixels*, com o esqueleto centralizado. Observe que cada junta tem uma cor própria.

Figura 38 – Exemplo de imagem SJoBB.

Fonte: Elaborada pelo autor.

4.2 Módulo de Reconhecimento

O módulo de reconhecimento é responsável pela identificação da ação que está sendo realizada no vídeo por meio das imagens SJoBB extraídas de cada frame do vídeo, conforme detalhado na Seção 4.1. A identificação da ação é realizada por um classificador previamente treinado para reconhecer a ação que está sendo executada, dentre um conjunto de ações previamente definidas.

A fim de utilizar as informações espaço-temporais presentes nos vídeos, propõe-se, como classificador, uma rede neural recorrente com camadas convolucionais. Na etapa de treinamento, o classificador aprende as características temporais da ação humana, analisando a sequência de quadros do vídeo de entrada por meio das propriedades de recorrência da rede neural recorrente ao mesmo tempo que aprende as características espaciais por meio das camadas de convolução 2D da rede neural.

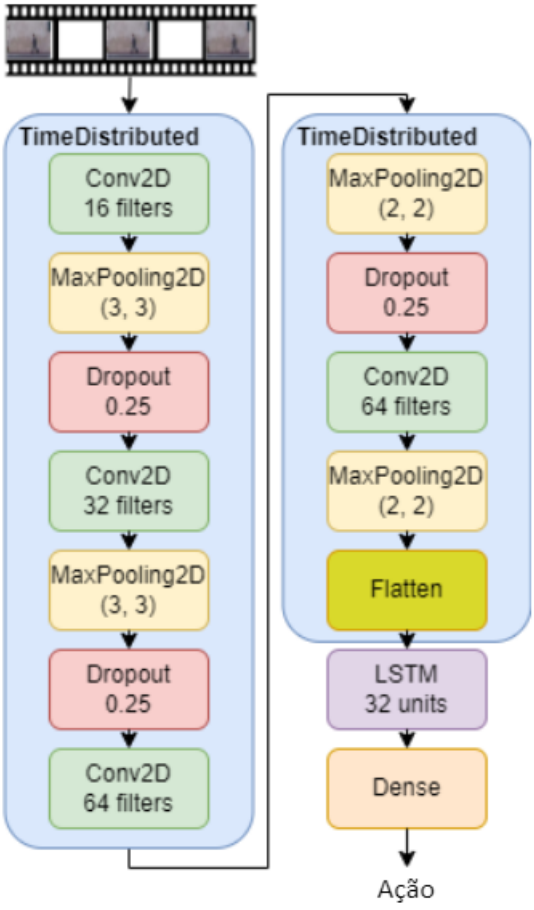
Todas as camadas da rede neural proposta, com exceção das LSTM, são encapsuladas por um camada *Time Distributed*, que garante que a mesma operação realizada pela camada que está sendo envolvida é aplicada a cada quadro de vídeo da mesma forma, com o mesmos pesos.

A arquitetura da rede neural adotada neste trabalho foi proposta por Donahue et al. (2015). Esta rede neural faz uso de uma LSTM logo após algumas camadas convolucionais. Todas as camadas de convolução possuem como função de ativação uma *Rectified Linear Unit* (ReLU) e um *kernel* de tamanho 3×3. No processo de treinamento da rede neural, uma função de *callback* foi configurada para monitorar o valor *validation loss* e restaurar os melhores pesos da rede ao final do treinamento. O modelo foi compilado utilizando um otimizador Adam e

treinado por 70 épocas. A Figura 39 apresenta a arquitetura da rede neural recorrente com camadas convolucionais adotada neste trabalho.

O método proposto nesta dissertação visa utilizar as imagens SJoBB como descritor das ações humanas em cada frame juntamente com o modelo de rede neural recorrente com camadas convolucionais proposto por Donahue et al. (2015) a fim de atingir desempenhos comparáveis aos trabalhos correlatos, estado da arte, para o problema de reconhecimento de ações humanas, utilizando um descritor simples e eficiente, bem como, técnicas modernas e eficientes de aprendizado de máquina.

Figura 39 – Arquitetura da rede neural recorrente com camadas convolucionais utilizada neste trabalho.



Fonte: Elaborada pelo autor.

5 Experimentos e Resultados

Além do método proposto nesta dissertação, apresentado na Seção 4, que utiliza as imagens SJoBB como descritores dos quadros dos vídeos e uma rede neural recorrente convolucional como classificador, outros descritores e classificadores foram avaliados a fim de se encontrar a melhor abordagem para o reconhecimento de ações humanas baseado nas articulações do esqueleto obtidas de poses 2D. Este capítulo descreve os experimentos realizados durante o desenvolvimento da pesquisa, bem como os resultados obtidos.

5.1 Bases de Dados e Recursos Utilizados

Os experimentos foram realizados utilizando o serviço do Google Colab Pro, fazendo uso de uma máquina com processador Intel Xeon CPU 2.30GHz, GPU Tesla P100 com 16GB e 26GB de memória RAM. Todo o desenvolvimento dos códigos foram feitos utilizando a linguagem de programação *Python* e suas bibliotecas, como por exemplo *Tensorflow* e *Scikit-Learn*.

Foram utilizadas as bases de dados KTH (SCHULDT; LAPTEV; CAPUTO, 2004) e Weizmann (GORELICK et al., 2007b), descritas nas Seções 3.3.1 e 3.3.2, respectivamente, por se tratarem de duas bases de dados amplamente utilizadas para a avaliação de métodos de reconhecimento de ações humanas e possuirem as características necessárias para o método proposto, como a presença de uma única pessoa nos vídeos e com o corpo inteiro do indivíduo sendo retratado durante a execução da ação.

Em todos os experimentos, os vídeos foram submetidos ao método PifPaf (KREISS; BERTONI; ALAHI, 2019) a fim de estimar as poses 2D em cada frame.

5.2 Experimento I - Imagens Médias

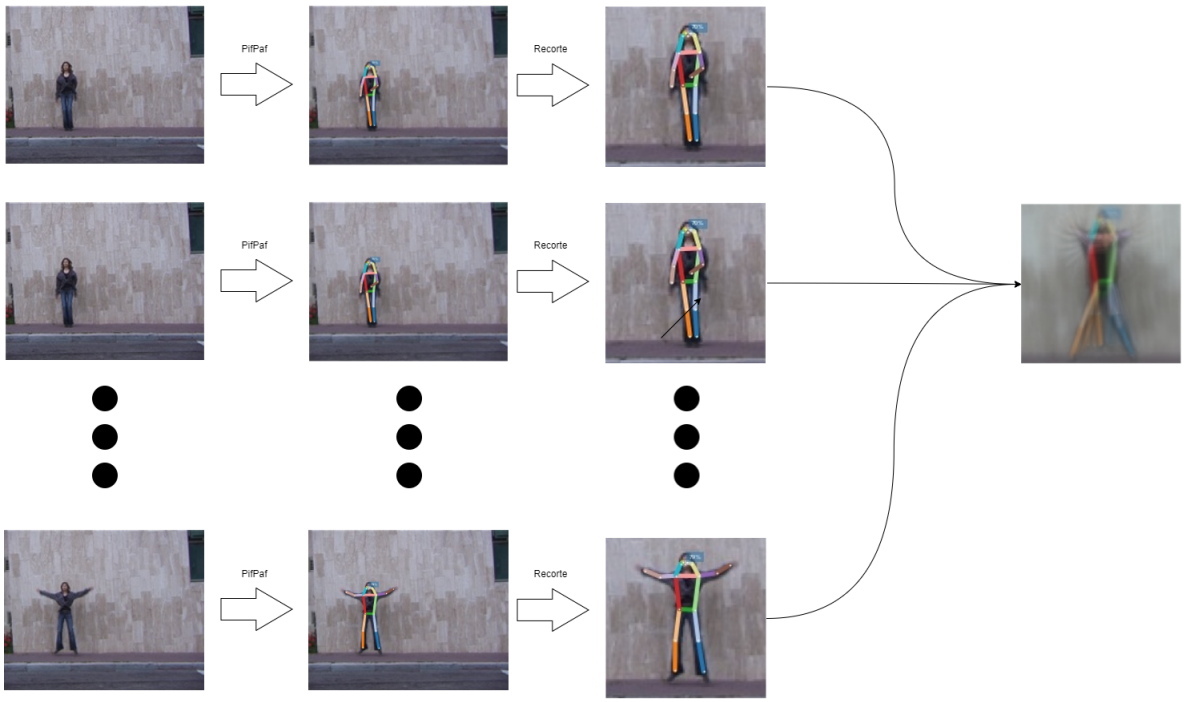
No início desta pesquisa, estava-se interessado em investigar descritores que pudessem representar de forma compacta o conteúdo de todo o vídeo contendo a ação humana. Inspirados em trabalhos que usam mapas de calor das juntas ou imagens médias das silhuetas humanas para o reconhecimento de ação humana, propôs-se a utilização de uma única imagem, gerada a partir das imagens criadas diretamente pelo método PifPaf, que contém o esqueleto estimado da pessoa sobreposto à imagem original, como pode ser visto na Figura 40.

Figura 40 – Exemplo de imagem do esqueleto gerada pelo método PifPaf.

Fonte: Elaborada pelo autor.

A Figura 41 ilustra o processo de geração da imagem média a partir das imagens do esqueleto gerado pelo método PifPaf e centralizadas para cada quadro do vídeo contendo a ação humana.

Figura 41 – Fluxo de geração de imagens médias dos esqueletos gerados pelo PifPaf a partir do vídeo de entrada.



Fonte: Elaborada pelo autor.

De forma análoga à estratégia utilizada para calcular as imagens médias a partir das imagens dos esqueletos geradas pelo PifPaf, foram obtidas imagens médias das imagens SJoBB, geradas a partir das coordenadas das juntas dos esqueletos, conforme descrito na Seção 4.1.

Com as imagens médias geradas, classificadores baseados em regressão logística e em CNN foram treinados utilizando a técnica de *leave-one-out* que consiste em treinar o classificador com todos os vídeos, deixando apenas um de fora para realizar o teste e repetindo-se esse processo para todas as imagens deixando sempre uma de fora do treino.

Para avaliar a eficácia dos classificadores utilizou-se a acurácia como métrica, uma vez que a maioria dos trabalhos correlatos encontrados na literatura, apresentados na Seção 3.2, fazem uso dessa métrica.

Para o classificador baseado em regressão logística, os parâmetros utilizados foram os padrões da biblioteca *Scikit-Learn* em Python. Para o classificador baseado em rede neural convolucional, adotou-se uma arquitetura simples, composta por duas camadas de convolução intercaladas por camadas de *max pooling* e uma camada densa no final.

A Tabela 2 mostra os resultados obtidos neste experimento para a base de dados Weizmann utilizando-se tanto as imagens médias derivadas das imagens dos esqueletos geradas diretamente pelo PifPaf, quanto as imagens médias derivadas das imagens SJoBB. Pode-se observar que as imagens SJoBB atingiram uma acurácia superior em relação às imagens dos esqueletos. Porém, comparando-se com as acurácias obtidas por outros métodos propostos na literatura na base de dados Weizmann, tanto a regressão logística quanto a CNN obtiveram resultados inferiores.

Tabela 2 – Acurácias obtidas pelos classificadores baseados em Regressão Logística e em Rede Neural Convolucional utilizando-se as imagens médias, para a base de dados Weizmann.

	Imagens médias SJoBB	**Imagens médias dos esqueletos**
Regressão Logística	**87.1%**	82.89%
CNN	**89.25%**	79.65%

5.3 Experimento II - Redes Neurais Recorrentes Convolucionais

Tendo em vista os resultados promissores obtidos com o uso das imagens JSoBB, decidiu-se continuar utilizando-as como descritores das ações humanas presentes nos quadros do vídeo. Entretanto, como os resultados estavam aquém daqueles obtidos por métodos correlatos estado-da-arte, decidiu-se explorar as potencialidades da análise temporal por meio de uma LSTM, associada à análise espacial por meio de uma CNN. Propôs-se, então, neste trabalho, o método descrito no Capítulo 4 para reconhecimento de ações humanas que combina imagens SJoBB com redes neurais recorrentes convolucionais.

Experimentos realizados com o classificador proposto, baseado em CNN e LSTM, evidenciaram ainda mais o potencial das imagens SJoBB. Nestes experimentos, foram utilizados quatro tipos diferentes de imagens para realizar a classificação das ações humanas: SJoBB

com cores distintas para cada junta, uma variante da SJoBB que utiliza uma única cor para todas as juntas, imagens das poses geradas diretamente pelo PifPaf e imagens originais obtidas diretamente dos quadros dos vídeos de entrada. A Tabela 3 mostra os resultados obtidos nestes experimentos. Pode-se observar que a rede neural recorrente convolucional em conjunto com os descritores SJoBB com cores distintas para cada junta apresenta acurácia superior aos demais descritores avaliados.

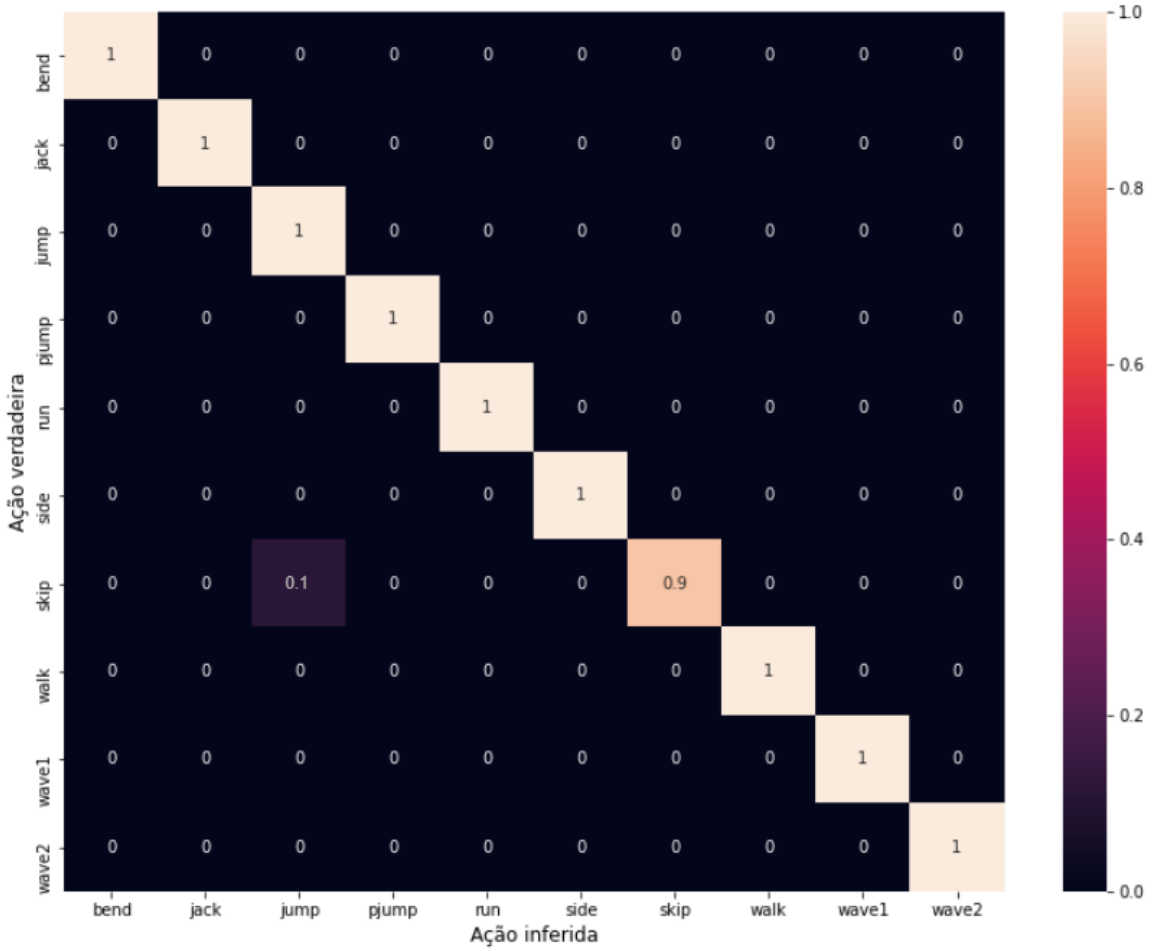
Tabela 3 – Acurácias obtidas pela rede neural recorrente convolucional para a base de dados Weizmann, utilizando-se quatro descritores distintos.

Descritor	**Acurácia**
SJoBB com Cores Distintas	**98.92%**
SJoBB com Cor Única	95.56%
Imagens das Poses PifPaf	90.32%
Imagens Originais	91.30%

É interessante notar os resultados obtidos pela utilização das imagens SJoBB onde todos os pontos das juntas foram desenhados no fundo preto contendo a mesma cor, diferente do método proposto originalmente, onde cada ponto possuiria uma única cor diferentes dos demais. Ao realizar essa alteração a acurácia teve uma queda de 98.92% para 95.56%, comprovando que o fato de cada junta ter uma única cor contribui para que a rede neural tenha um melhor aprendizado das ações humanas. Além disso, os resultados obtidos evidenciam a superioridade do descritor SJoBB em comparação ao descritor representado pelas imagens das poses 2D e também ao uso das imagens diretamente obtidas dos frames.

A Figura 42 mostra a matriz de confusão para a base de dados Weizmann referente a acurácia obtida de 98.92%. O único erro que o método cometeu foi ao confundir a ação *skip* com *jump*, onde a primeira é pular com apenas uma perna e a segunda com as duas pernas.

Figura 42 – Matriz de confusão normalizada base de dados Weizmann.



Fonte: Elaborada pelo autor.

5.4 Experimento III - Mecanismos de Atenção

Outra hipótese que foi avaliada neste trabalho foi a de que a utilização de mecanismos de atenção poderia melhorar ainda mais a acurácia apresentada pela rede neural recorrente convolucional com os descritores SJoBB.

A Tabela 4 mostra as acurácias obtidas para as bases de dados Weizmann e KTH para o método original e suas variantes nos quais foram inseridas camadas de atenção espacial e camadas de atenção temporal. Observa-se que os mecanismos de atenção, contrariando as expectativas, pioraram a acurácia obtida pelo método original. Uma possível explicação para este resultado, tendo em vista os benefícios geralmente propiciados pelos mecanismos de atenção, é que o descritor SJoBB já incorpora, intrinsicamente, as informações mais relevantes para o reconhecimento de ações humanas baseadas nos esqueletos obtidos de poses 2D. Esta tabela mostra também que o mecanismo de atenção temporal propicia uma maior acurácia do

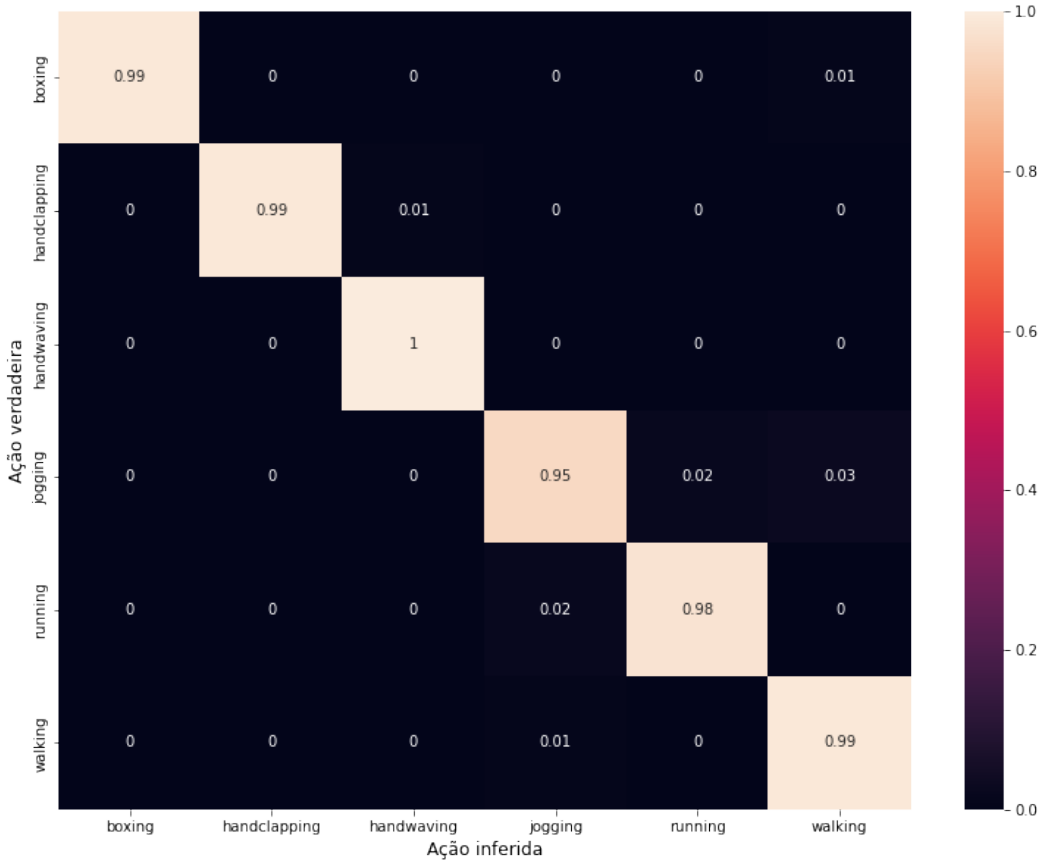
que o mecanismo de atenção espacial.

Tabela 4 – Acurácias obtidas pela rede neural com a arquitetura original e pela rede neural utilizando mecanismos de atenção.

Arquitetura da Rede Neural	**Weizmann**	**KTH**
Original	**98.92%**	**98.33%**
Original + Atenção Espacial	91.39%	96.66%
Original + Atenção Temporal	93.54%	95.82%

A matriz de confusão apresentada na Figura 43 mostra quais erros foram cometidos pelo método para a base KTH, onde obteve-se 98.33% de acurácia. Observa-se que as ações *jogging*, *running* e *walking* foram as que apresentaram erros, o que é compreensível pois trotar, correr e andar são ações muito parecidas.

Figura 43 – Matriz de confusão normalizada base de dados KTH.



Fonte: Elaborada pelo autor.

5.5 Comparação com Outros Métodos

O método proposto neste trabalho para o reconhecimento de ações humanas baseadas no descritor SJoBB e em redes neurais recorrentes convolucionais foi comparado com catorze métodos correlatos, propostos nos últimos dez anos e descritos na Seção 3.2.

A Tabela 5 mostra as acurácias obtidas pelo método proposto neste trabalho, bem como pelos outros catorze trabalhos correlatos quando aplicados nas bases de dados Weizmann e KTH. Observa-se nesta tabela que o método proposto obteve a terceira maior acurácia para a base de dados Weizmann e a segunda maior acurácia para a base de dados KTH.

Tabela 5 – Comparação das acurácias (%) obtidas pelo método proposto e outros métodos encontrados na literatura, para as bases de dados Weizmann e KTH.

Método	**Ano**	**Weizmann**	**KTH**
Método Proposto (Rede Neural Recorrente Convolunional com SJoBB)	2022	**98.92**	**98.33**
(GUO; ISHWAR; KONRAD, 2013)	2013	100.00	98.50
(ALCANTARA et al., 2017)	2017	100.00	92.20
(CARMONA; CLIMENT, 2018)	2018	98.80	97.50
(MOREIRA; MENOTTI; PEDRINI, 2020)	2020	98.00	97.50
(SILVA, 2020)	2020	97.85	97.16
(SINGH; VISHWAKARMA, 2019)	2019	97.66	94.50
(CHOU et al., 2018)	2018	95.56	90.58
(ZHANG; TAO, 2012)	2012	93.87	93.50
(CHAARAOUI; CLIMENT-PÉREZ; FLÓREZ-REVUELTA, 2013)	2013	90.32	-
(JUNEJO; AGHBARI, 2012)	2012	88.60	-
(RAVANBAKHSH et al., 2015)	2015	-	95.60
(DOUMANOGLOU; VRETOS; DARAS, 2016)	2016	-	88.70
(JI et al., 2012)	2012	-	90.20
(ALMEIDA et al., 2017)	2017	-	96.80

6 Conclusões

Conforme a capacidade de processamento e armazenamento de grandes quantidades de dados avança com as tecnologias atuais, o tópico de reconhecimento de imagens ganha mais espaço e se torna, cada vez mais, um tema de pesquisa relevante, tanto para monitorar áreas públicas em que a principal motivação atual se dá em questões de segurança, como para residências privadas para monitoramento e rápido atendimento de possíveis acidentes domésticos que possam ocorrer.

Com o surgimento de métodos robustos, eficazes e eficientes para estimação de poses 2D, como o OpenPose (CAO et al., 2019) e o PifPaf (KREISS; BERTONI; ALAHI, 2019), um novo leque de possibilidades para o reconhecimento de ações humanas se abriu, surgindo diferentes abordagens para resolver esse problema, como por exemplo no estudo de Silva e Marana (2020) o qual gerou novas *features* a partir dos resultados da estimação de pose, calculando os ângulos e as trajetórias das juntas para realizar a classificação.

As poses 2D têm como um dos seus principais benefícios a possibilidade de preservar a privacidade das pessoas envolvidas na cena, uma vez que, seja no caso das imagens geradas dos esqueletos das poses do PifPaf ou mesmo as imagens das localizações das juntas, as informações que poderiam ser usadas para identificar uma pessoa não estão presentes na imagem, fazendo com que soluções de mercado possam se beneficiar dessa técnica e tenham argumentos para incluírem suas soluções no dia a dia das pessoas sem que elas se preocupem em terem sua privacidade exposta.

Paralelo a isso, as redes neurais profundas e o aprendizado profundo vem demonstrando ótimos resultados em diferentes áreas que possam se beneficiar com essa tecnologia, seja para problemas que envolvam visão computacional ou qualquer outro tipo de problema. Particularmente para o caso do reconhecimento de ações humanas, por se utilizar imagens como dados de entrada, o uso de CNNs deve ser levado em consideração pela natureza dessa arquitetura em resolver esse tipo de problema. Somado a isso, RNNs também são relevantes para a solução do problema, uma vez que, por se tratar de uma sequência de *frames* em que a informação temporal é importante para a classificação, a RNN consegue, através das LSTM, captar a ação que a pessoa esta realizando ao longo do tempo.

Outra conclusão relevante diz respeito à importância de ambas as informações temporais e espaciais para o papel do reconhecimento de ações humanas. Nos experimentos realizados, quando utilizadas as arquiteturas e algoritmos que trabalhavam apenas com as informações espaciais, os resultados foram muito inferiores comparados àqueles obtidos quando utilizaram-se as dimensões espaço-tempo em conjunto.

Por fim, com relação aos descritores propostos e avaliados neste trabalho, baseados nas articulações dos esqueletos obtidos das poses 2D, concluiu-se que o uso de cores distintas para representar cada articulação contribui significativamente para aumentar a acurácia do reconhecimento das ações, pois quando foi utilizada a mesma cor para todos os pontos, a acurácia teve uma queda de 3.36%, o que mostra a importância da informação visual detalhada para o reconhecimento da ação.

6.1 Direcionamentos para Trabalhos Futuros

Para trabalhos futuros, alguns pontos podem ser explorados:

- Utilização do OpenPose para obtenção das poses 2D e comparação com o PifPaf;
- Avaliação de outras arquiteturas para a rede neural utilizada no módulo de reconhecimento;
- Avaliar outras implementações de mecanismos de atenção para integrar à arquitetura;
- Limitar o número de juntas buscando identificar aquelas que podem melhor representar as ações humanas;
- Aplicar o método proposto em outras bases de dados de ações humanas e avaliar o seu desempenho.

6.2 Publicação Realizada

Resultados parciais desta dissertação de mestrado foram descritos em um artigo científico que foi apresentado no *11th Brazilian Conference on Intelligent Systems (BRACIS 2022)* e publicado na Lecture Notes in Computer Science - Springer.

1. Belluzzo, B., Marana, A.N. (2022). Human Action Recognition Based on 2D Poses and Skeleton Joints. In: Xavier-Junior, J.C., Rios, R.A. (eds) Intelligent Systems. BRACIS 2022. Lecture Notes in Computer Science(), vol 13654 . Springer, Cham. `$https://doi.org/10.1007/978-3-031-21689-3_6$` (BELLUZZO; MARANA, 2022)

Referências

AGGARWAL, J. K.; RYOO, M. S. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, Acm New York, NY, USA, v. 43, n. 3, p. 1–43, 2011.

ALBAWI, S. et al. Social touch gesture recognition using convolutional neural network. *Computational Intelligence and Neuroscience*, Hindawi, v. 2018, 2018.

ALCANTARA, M. F. de et al. Action identification using a descriptor with autonomous fragments in a multilevel prediction scheme. *Signal, image and video processing*, Springer, v. 11, n. 2, p. 325–332, 2017.

ALMEIDA, R. et al. Human action classification using an extended bow formalism. In: SPRINGER. *International Conference on Image Analysis and Processing*. [S.l.], 2017. p. 185–196.

ANTIPOV, G. et al. Learned vs. hand-crafted features for pedestrian gender recognition. In: *Proceedings of the 23rd ACM international conference on Multimedia*. [S.l.: s.n.], 2015. p. 1263–1266.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

BELLUZZO, B.; MARANA, A. N. Human action recognition based on 2d poses and skeleton joints. In: XAVIER-JUNIOR, J. C.; RIOS, R. A. (Ed.). *Intelligent Systems*. Cham: Springer International Publishing, 2022. p. 71–83. ISBN 978-3-031-21689-3.

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, IEEE, v. 5, n. 2, p. 157–166, 1994.

CAO, Z. et al. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 43, n. 1, p. 172–186, 2019.

CARMONA, J. M.; CLIMENT, J. Human action recognition by means of subtensor projections and dense trajectories. *Pattern Recognition*, Elsevier, v. 81, p. 443–455, 2018.

CARREIRA, J.; ZISSERMAN, A. Quo vadis, action recognition? a new model and the kinetics dataset. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2017. p. 6299–6308.

CHAARAOUI, A. A.; CLIMENT-PÉREZ, P.; FLÓREZ-REVUELTA, F. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, Elsevier, v. 34, n. 15, p. 1799–1807, 2013.

CHENG, G. et al. Advances in human action recognition: A survey. *arXiv preprint arXiv:1501.05964*, 2015.

CHOU, K.-P. et al. Robust feature-based automated multi-view human action recognition system. *IEEE Access*, IEEE, v. 6, p. 15283–15296, 2018.

DOLLÁR, P. et al. Behavior recognition via sparse spatio-temporal features. In: IEEE. *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. [S.l.], 2005. p. 65–72.

DONAHUE, J. et al. Long-term recurrent convolutional networks for visual recognition and description. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 2625–2634.

DOUMANOGLOU, A.; VRETOS, N.; DARAS, P. Action recognition from videos using sparse trajectories. IET, 2016.

DU, W.; WANG, Y.; QIAO, Y. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2017. p. 3725–3734.

DUARTE, G. *Redes Neurais | Redes Neurais Recorrentes*. 2019. Disponível em: <https://medium.com/turing-talks/turing-talks-26-modelos-de-prediÃğÃčo-redes-neurais-recorrentes-439198e9ecf3>. Acesso em: 19 jul. 2020.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016.

GORELICK, L. et al. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 29, n. 12, p. 2247–2253, 2007.

GORELICK, L. et al. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, v. 29, n. 12, p. 2247–2253, December 2007.

GRAHAM, B. Fractional max-pooling. *arXiv preprint arXiv:1412.6071*, 2014.

GUO, K.; ISHWAR, P.; KONRAD, J. Action recognition from video using feature covariance matrices. *IEEE Transactions on Image Processing*, IEEE, v. 22, n. 6, p. 2479–2494, 2013.

GUO, M.-H. et al. Attention mechanisms in computer vision: A survey. *arXiv preprint arXiv:2111.07624*, 2021.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

IBRAHIM, M. S. et al. A hierarchical deep temporal model for group activity recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2016. p. 1971–1980.

INSAFUTDINOV, E. et al. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2016. p. 34–50.

JHUANG, H. et al. Towards understanding action recognition. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2013. p. 3192–3199.

JI, S. et al. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 35, n. 1, p. 221–231, 2012.

JUNEJO, I. N.; AGHBARI, Z. A. Using sax representation for human action recognition. *Journal of Visual Communication and Image Representation*, Elsevier, v. 23, n. 6, p. 853–861, 2012.

JUNIOR. *Redes Neurais Recorrentes — LSTM*. 2019. Disponível em: <https://medium.com/@web2ajax/redes-neurais-recorrentes-lstm-b90b720dc3f6>. Acesso em: 21 dez. 2022.

KARPATHY, A. et al. Large-scale video classification with convolutional neural networks. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1725–1732.

KONG, Y.; FU, Y. Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*, 2018.

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

LAPTEV, I. On space-time interest points. *International journal of computer vision*, Springer, v. 64, n. 2-3, p. 107–123, 2005.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Ieee, v. 86, n. 11, p. 2278–2324, 1998.

LI, Z. et al. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, Elsevier, v. 166, p. 41–50, 2018.

LIN, M.; CHEN, Q.; YAN, S. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

LIN, T.-Y. et al. Microsoft coco: Common objects in context. In: SPRINGER. *European conference on computer vision*. [S.l.], 2014. p. 740–755.

LIPTON, Z. C.; BERKOWITZ, J.; ELKAN, C. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

M. MAITHANI. *Guide to OpenPose for Real-time Human Pose Estimation*. 2021. Disponível em: <https://analyticsindiamag.com/guide-to-openpose-for-real-time-human-pose-estimation/>. Acesso em: 6 abr. 2021.

MESSING, R.; PAL, C.; KAUTZ, H. Activity recognition using the velocity histories of tracked keypoints. In: IEEE. *2009 IEEE 12th international conference on computer vision*. [S.l.], 2009. p. 104–111.

MOREIRA, T. P.; MENOTTI, D.; PEDRINI, H. Video action recognition based on visual rhythm representation. *Journal of Visual Communication and Image Representation*, Elsevier, v. 71, p. 102771, 2020.

O'SHEA, K.; NASH, R. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

PAREEK, P.; THAKKAR, A. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, Springer, v. 54, n. 3, p. 2259–2322, 2021.

PARSA, B. *Deep Learning Methods for Video-Based Human Activity Recognition in Industrial Settings*. [S.l.]: University of Washington, 2020.

RANGASAMY, K. et al. Deep learning in sport video analysis: a review. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, v. 18, n. 4, p. 1926–1933, 2020.

RAVANBAKHSH, M. et al. Action recognition with image based cnn features. *arXiv preprint arXiv:1512.03980*, 2015.

ROHRER, B. *Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)*. 2017. Disponível em: <https://www.youtube.com/watch?v=WCUNPb-5EYIt=1348s>. Acesso em: 22 dez. 2022.

RYOO, M. S.; AGGARWAL, J. K. Semantic representation and recognition of continued and recursive human activities. *International journal of computer vision*, Springer, v. 82, n. 1, p. 1–24, 2009.

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

SHARMA, S.; KIROS, R.; SALAKHUTDINOV, R. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.

SILVA, M. V. d. Human action recognition based on spatiotemporal features from videos. Universidade Federal de São Carlos, 2020.

SILVA, M. V. da; MARANA, A. N. Human action recognition in videos based on spatiotemporal features and bag-of-poses. *Applied Soft Computing*, Elsevier, v. 95, p. 106513, 2020.

SIMONYAN, K.; ZISSERMAN, A. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

SINGH, S.; VELASTIN, S. A.; RAGHEB, H. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In: IEEE. *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*. [S.l.], 2010. p. 48–55.

SINGH, T.; VISHWAKARMA, D. K. A hybrid framework for action recognition in low-quality video sequences. *arXiv preprint arXiv:1903.04090*, 2019.

SOOMRO, K.; ZAMIR, A. R. Action recognition in realistic sports videos. In: *Computer vision in sports*. [S.l.]: Springer, 2014. p. 181–208.

SUN, Z. et al. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–20, 2022.

TOSHEV, A.; SZEGEDY, C. Deeppose: Human pose estimation via deep neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2014. p. 1653–1660.

VASWANI, A. et al. *Attention Is All You Need*. arXiv, 2017. Disponível em: <https://arxiv.org/abs/1706.03762>.

VIJAYAPRABAKARAN, K.; SATHIYAMURTHY, K.; PONNIAMMA, M. Video-based human activity recognition for elderly using convolutional neural network. *International Journal of Security and Privacy in Pervasive Computing (IJSPPC)*, IGI Global, v. 12, n. 1, p. 36–48, 2020.

VRIGKAS, M.; NIKOU, C.; KAKADIARIS, I. A. A review of human activity recognition methods. *Frontiers in Robotics and AI*, Frontiers, v. 2, p. 28, 2015.

WANG, H. et al. Action recognition by dense trajectories, 2011. *CrossRef][Google Scholar]*, p. 3169–3176, 2011.

WANG, H.; SCHMID, C. Action recognition with improved trajectories. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2013. p. 3551–3558.

WANG, H. et al. Evaluation of local spatio-temporal features for action recognition. In: BMVA PRESS. *Bmvc 2009-british machine vision conference*. [S.l.], 2009. p. 124–1.

WANG, S.-H. et al. Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling. *Frontiers in neuroscience*, Frontiers, v. 12, p. 818, 2018.

WEINLAND, D.; RONFARD, R.; BOYER, E. Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding*, Elsevier, v. 104, n. 2-3, p. 249–257, 2006.

WOLF, C. et al. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, Elsevier, v. 127, p. 14–30, 2014.

WU, J. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, v. 5, n. 23, p. 495, 2017.

XIA, L.; CHEN, C.; AGGARWAL, J. 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.], 2012. p. 20–27.

YU, G.; YUAN, J.; LIU, Z. Unsupervised random forest indexing for fast action search. In: IEEE. *CVPR 2011*. [S.l.], 2011. p. 865–872.

ZHANG, Z.; TAO, D. Slow feature analysis for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 34, n. 3, p. 436–450, 2012.

APÊNDICE A – *Código Fonte*

Durante o desenvolvimento desta dissertação de mestrado o código fonte foi disponibilizado em um repositório no GitHub contendo os notebooks com todas as técnicas utilizadas durante o processo de extração de características e experimentação de diferentes tipos de algoritmos para realizar a classificação da ação humana. O código pode ser acessado em:

<https://github.com/BruBel/SJoBBHumanPoseEstimation>