Vitor Thinassi Basilio

# Reconhecimento de Ações usando RNA

Viçosa, MG

2020

Vitor Thinassi Basilio

# Reconhecimento de Ações usando RNA

Monografia apresentada ao Departamento de Engenharia Elétrica do Centro de Ciências Exatas e Tecnológicas da Universidade Federal de Viçosa, para a obtenção dos créditos da disciplina ELT 402 – Projeto de Engenharia 2 – e cumprimento do requisito parcial para obtenção do grau de Bacharel em Engenharia Elétrica.

Orientador Prof. Dr. Alexandre Santos Brandão

Coorientador: Ms. Kevin Braathen de Carvalho

Viçosa, MG

2020

Vitor Thinassi Basilio

# Reconhecimento de Ações usando RNA

Monografia apresentada ao Departamento de Engenharia Elétrica do Centro de Ciências Exatas e Tecnológicas da Universidade Federal de Viçosa, para a obtenção dos créditos da disciplina ELT 402 – Projeto de Engenharia 2 – e cumprimento do requisito parcial para obtenção do grau de Bacharel em Engenharia Elétrica.

Trabalho aprovado em 08 de dezembro de 2020.

COMISSÃO EXAMINADORA

**Prof. Dr. Alexandre Santos Brandão**
Orientador

**Ms. Kevin Braathen de Carvalho**
Coorientador

**Prof. Dra. Ketia Soares Moreira**
Arguidora 1

**Prof. Dr. Rodolpho Vilela Alves Neves**
Arguidor 2

Viçosa, MG
2020

*Esta monografia é dedicada a todas as pessoas que lutam por um mundo melhor, por mais igualdade e respeito.*

# Agradecimentos

Agradeço ao apoio e ao ensinamento das pessoas que, de alguma forma, me ensinaram e fizeram refletir, que me ajudaram a transformar um simples passo em combustível para que minha jornada fosse tranquila e com poucas turbulências.

Agradeço especialmente à minha família, que sempre me apoiou. Aos meus pais, José Ricardo Basílio e Elineia Thinassi Basílio, pelo amor e carinho que têm por mim, por terem paciência para entender meus problemas, por estarem sempre lutando pela minha felicidade e de meu irmão.

Ao meu irmão, por me aturar vinte e quatro anos e, mesmo nunca tendo deixado eu ser o *player* 1, me ensinou a ter mais cuidado com as palavras e ter consciência de minha fala.

Às minhas primas e primos por parte de pai, que sempre estão me fazendo sorrir, companheiras de lanches, de piscina, caminhadas e gritarias.

Às minhas primas e primos por parte de mãe, exemplos a serem seguidos, sempre presentes nos almoços de domingo na casa do Vô Juju.

Ao Lucas, meu primo, que morou comigo durante quatro anos em Viçosa. Teve paciência para aguentar minha agitação dentro de casa, sempre querendo fazer algo diferente. Foi meu companheiro de jogos online, junto com nosso outro primo, Jhonathan.

Às minhas tias, que considero como mães. Sei que sempre posso contar com elas, exemplos de mulheres batalhadoras e independentes.

Aos meus tios, que também considero como pais. Estão sempre dispostos a conversas longas, com boas histórias e ensinamentos.

Ao Vô Juju e à Vó Conceição, que mesmo sem saber lidar com as grandes mudanças, na sua simplicidade sempre demonstraram amor e carinho por mim e por toda família. À Vó Cornelha e ao Vô Zeca, que, infelizmente, conheci pouco, mas sei que me protegem e ajudam com forças espirituais.

Ao Júnior, por me aguentar durante dois anos, tirando minhas dúvidas de português e de todas as outras línguas, por me ensinar que coletividade também está em pequenas ações.

Aos meus amigos de Ubá, do ensino médio e alguns do ensino fundamental, que foram uma fortaleza durante os anos de escola. É muito bom poder contar com essas amizades, que duram até hoje.

Aos meus amigos do VELT, pela luta constante por um mundo com mais respeito

às minorias e às diversidades.

A todos meus amigos da ELT15, em especial ao Ricardo Ferreira, Gabriel Rabelo e Lucas Jonys, companheiros de batalhas e estudos durante esses longos anos de curso. E aos Farofas, amigos com quem eu pude contar em todos instantes.

Aos meus colegas do CAELT, em especial à Patricia Pontes, por sempre ter bons conselhos e propiciar ótimas trocas durante nossa jornada.

À Juliana e Maria Teresa, por serem minhas melhores companheiras de viagens e ideias malucas.

Aos demais amigos que a UFV e Viçosa me deram, pelos momentos de que sempre lembrarei em minha vida.

Aos meus colegas de quarto Juan, Cristina e Ian, que, durante seis meses, foram como uma família para mim na Espanha. Acolheram-me e sempre me deram os melhores conselhos e dicas.

Aos meus amigos do intercâmbio, por todo companheirismo, e que, à primeira vista, eu soube que eram pessoas com quem eu poderia contar para o que der e vier.

Ao meu co-orientador e amigo de pesquisa Kevin Braathen, pelas horas quebrando a cabeça desenvolvendo o trabalho que deu origem a esta monografia. Pelos momentos de frustração e erros, que, com calma, conseguimos superar e aprender com eles ainda mais.

Ao meu orientador Alexandre Santos Brandão. Sem ele um simples passo na minha jornada não seria possível. Obrigado por abrir portas para mim e para muitos outros alunos.

Ao CNPq e à FAPEMIG, agências de fomento que me concederam bolsas de iniciação científica.

Aos meus professores do Departamento de Engenharia Elétrica e a todos aqueles que dedicaram suas carreiras ao ensino e à pesquisa. Sem eles nada seria possível.

A todos vocês, meu muito obrigado,
Vitor Thinassi Basilio.

“Uma jornada de mil quilômetros precisa começar com um simples passo“
(Provérbio Chinês)
“Uma busca começa sempre com a Sorte de Principiante.
E termina sempre com a Prova do Conquistador“
(*O Alquimista*. Paulo Coelho)

# Resumo

O reconhecimento de ação vem ganhando espaço em pesquisas em razão das suas mais variadas aplicações. Esta monografia propõe um método de reconhecimento de ações usando Rede Neural Artificial (RNA), baseado na redução da dimensão das entradas do sistema, cujo intuito é necessitar de uma base de dados pequena para treino. O método foi desenvolvido para ser aplicado a interações com robôs, segundo a Robótica Socialmente Assistiva (SAR). Onze classes foram escolhidas, nas quais nove são ações padrões e duas são ações neutras. A base de dados é criada por cinco pessoas com estilos diferentes de corpo, para fins de generalização. Testes *offline* e *online* foram realizados para demonstrar e validar a precisão do método proposto. Os testes *offline* foram divididos em três, de forma a testar como o método se comporta com números variados de amostras na entrada. Os testes *online* foram divididos em dois: o primeiro serviu para validar o método, no qual, ao mesmo tempo que as ações são feitas por um usuário, elas são classificadas; o segundo foi uma simulação de aplicação da vida real, feita através de um Jogo da Velha, em que um robô é controlado pelas ações feitas pelos usuários. Os resultados foram expostos em matrizes de confusão para esclarecê-los; eles mostram que o método obteve uma alta taxa de acerto, acima de $97,5\%$ em todos os testes. Portanto, o algoritmo proposto é capaz de identificar padrões de ações, até mesmo para uma base de dados pequena - com 5 a 10 amostras coletadas de cada pessoa para cada classe. É notável como o método permite incorporar novos usuários com facilidade, uma vez que demanda um baixo número de repetições, para que as amostras possam ser adicionadas à base de dados e um novo usuário possa ser reconhecido. Além disso, os resultados demonstram a possibilidade da utilização desse método em diversas aplicações da vida real.

**Palavras-chaves**: Robótica; *Microsoft Kinect*; Rede Neural Artificial; Reconhecimento de Ações.

# Abstract

Action recognition has been gaining interest in research due to its great number of applications. This work proposes an action recognition method using Artificial Neural Network (ANN), based on dimension reduction off the system's inputs in order to require a small database for training. The method was developed to be used on Social Assistive Robotics (SAR) or humam interaction with robots. Eleven classes were chosen, in which nine are standard actions and the other two are neutral actions. The dataset was created by five people with different body shape, for generalization and robustness purposes. Offline and Online tests were performed to demonstrate and validate the accuracy of the proposed method. The offline tests were divided into three, in order to test how the method behaves with varying numbers of samples at the entrance and the number of skeletons. The online tests were divided into two: the first was to validate the method, in which, at the same time as the actions are performed by a user, they are classified; the second was a simulation of real life application, made through a TicTacToe game, in which a robot is controlled by the actions taken by the users, that represents a spot in the game. The results were exposed in confusion matrices to clarify them; they show that the method achieved a high accuracy, above 97.5% in all tests. Therefore, the proposed algorithm is able to identify action patterns, even for a small database - with 5 to 10 samples collected from each person for each class. The method allows to incorporate new users with ease, since it requires a low number of repetitions, so that the samples can be added to the dataset and a new user can be classified. In addition, the results demonstrate the possibility of using this method in several real-life applications.

**Key-words**: Robotics; Microsoft Kinect; Artificial Neural Network; Action Recognition.

# Lista de ilustrações

Figura 1 – Exemplos da interface entre humano e robô. . . . . 14
Figura 2 – Esqueleto com as 25 juntas destacadas em verde . . . . . 18
Figura 3 – Exemplos de ações encontradas em base de dados da internet. . . . . 20
Figura 4 – Ações padrões da base de dados. . . . . 22
Figura 5 – Juntas do esqueleto destacadas para aquelas utilizadas na classificação 23
Figura 6 – Representação da RNA utilizada . . . . . 25
Figura 7 – Fluxograma para o algoritmo de reconhecimento de gestos . . . . . 25
Figura 8 – Imagens para sinalizar qual ação deve ser realizada . . . . . 26
Figura 9 – Posições referentes às ações para o Jogo da Velha. . . . . 26
Figura 10 – Ambiente criado para simulação de aplicação. . . . . 32

# Lista de tabelas

Tabela 1 – Especificação do sensor *Kinect v2.0* . . . . . 19
Tabela 2 – Ensaios realizados para teste e validação do método proposto . . . . . 27
Tabela 3 – Matriz de confusão para o teste 2P-10A: 98,3% taxa de acerto . . . . . 29
Tabela 4 – Matriz de confusão para o teste 5P-10A: 99,4% taxa de acerto . . . . . 29
Tabela 5 – Matriz de confusão para o teste 5P-05A: 97,5% taxa de acerto . . . . . 30
Tabela 6 – Matriz de confusão para o primeiro teste *online*: 98% taxa de acerto . 31

# Lista de abreviaturas e siglas

| | |
|---|---|
| RNA | Rede Neural Artificial |
| SAR | Robótica Socialmente Assistiva |
| SIR | Robótica Social Interativa |
| AR | Robótica Assistiva |
| RGB | *Red-Green-Blue* |
| RGB-D | *Red-Green-Blue Depth* |
| DTW | *Dynamic Time Warping* |
| HMM | *Hidden Markov Model* |
| SDK | *Software Development Kit* |
| FPS | Quadros por segundo |

# Sumário


**1 INTRODUÇÃO . . . . . 13**
**1.1 Reconhecimento de Ações . . . . . 15**
**1.2 Objetivos e Estrutura . . . . . 16**

**2 MATERIAIS E MÉTODOS . . . . . 18**
**2.1 Sensor de Movimento . . . . . 18**
**2.2 Base de dados . . . . . 19**
**2.3 Classificação . . . . . 23**
2.3.1 Classificação *Offline* . . . . . 24
2.3.2 Classificação *Online* . . . . . 24

**3 RESULTADOS E DISCUSSÃO . . . . . 27**
**3.1 Classificação *Offline* . . . . . 28**
3.1.1 Teste 2P-10A . . . . . 28
3.1.2 Teste 5P-10A . . . . . 29
3.1.3 Teste 5P-05A . . . . . 30
**3.2 Classificação *Online* . . . . . 31**
3.2.1 Primeiro Teste . . . . . 31
3.2.2 Segundo Teste . . . . . 32

**4 CONSIDERAÇÕES FINAIS . . . . . 34**

**REFERÊNCIAS . . . . . 36**

# 1 Introdução

Esta monografia é um dos resultados do projeto de Iniciação Cientifica "Rastreamento de Esqueleto e Reconhecimento de Gestos Utilizando o Sensor *Kinect*", desenvolvido no Núcleo de Especialização em Robótica (NERO). Baseou-se também no artigo, publicado no Simpósio Brasileiro de Automação Inteligente (SBAI) - 2019, intitulado "Reconhecimento de Ações por RNA em Aplicações de Robótica Social" (BASILIO; CARVALHO; BRANDãO, 2019); e em outro, sob revisão, com o título de *"Action Recognition for Educational Proposals applying Concepts of Social Assistive Robotic"*.

Robótica Socialmente Assistiva (*Social Assistive Robotic*, SAR) pode ser definida como a interseção entre Robótica Social Interativa (*Social Interactive Robotics*, SIR) e Robótica Assistiva (*Assistive Robotics*, AR). SIR, inicialmente definida por Fong, Nourbakhsh e Dautenhahn (2003) e depois detalhada por Feil-Seifer e Mataric (2005). É inspirada pela comunicação entre robôs e o ambiente, podendo ocorrer também entre as próprias máquinas. Em outras palavras, quando há algum tipo de interação, os conceitos de SIR são aplicados. Já AR ocorre quando uma máquina auxilia um ser humano (FEIL-SEIFER; MATARIC, 2005).

Diversos trabalhos têm SAR como campo de estudo, sendo aplicada a projetos de auxílio a pessoas idosas (TAPUS; MAJA; SCASSELLATTI, 2007), indivíduos em reabilitação (KAHN et al., 2001) ou até mesmo para robótica educacional (BENITTI, 2012). Assim, a interface entre humano e robô pode ser realizada de diferentes maneiras, como:

1. **Gesto ou ação**, a linguagem corporal pode ser útil em situações em que tanto o humano quanto o robô interagem com o ambiente. Além disso, quando utilizada para terapias físicas, o reconhecimento da posição e orientação do corpo é vital e pode reafirmar a comunicação entre o usuário e o agente (KANDA et al., 2004).

2. **Comando de voz**, utilizado no dia a dia na comunicação humana, também podendo ser aplicado para dialogar com um robô, especialmente com aqueles que usam gerador de comandos de voz sintéticos ou voz humana pré-gravada (ERIKSSON, 2004).

3. **Periféricos**, como *mouse*, teclado e tela *touchscreen*, podem ser úteis em situações nas quais o reconhecimento de voz e gesto sejam tediosos, como apontar para um local no mapa. Desta maneira, utilizar dispositivos periféricos pode ser mais eficiente e natural para alguns usuários (HUTTENRAUCH; EKLUNDH, 2002; MONTEMERLO et al., 2002).

A Figura 1 representa algumas dessas maneiras de comunicação.

Figura 1 – Exemplos da interface entre humano e robô.

(a) Gesto ou Ação.


(b) Comando de voz.

(c) Periféricos.

Fontes: 1a - Bruno Oliveira/G1; 1b - Freepick; 1c - Freepick.

Com a atual pandemia do COVID-19, diversos estudos mostram que um dos meios de transmissão do vírus se dá pelo contato com superfícies contaminadas (PASCARELLA et al., 2020). Portanto, métodos de interação entre seres humanos e robôs, que evitem o contato direto com superfícies, como Comandos de Voz e Gestos ou Ações, podem ser úteis em diversas aplicações. Esse é o caso do método proposto com a utilização de gestos/ações para comunicação e controle, em possíveis aplicações em Robótica Socialmente Assistiva.

Na literatura, os termos ação e gesto não se encontram bem definidos. Ação em Heidari e Iosifidis (2020) é usado para definir os movimentos realizados com o corpo inteiro; já Li et al. (2020), associa ação a qualquer tipo de detecção de movimentos gerais e específicos do corpo humano, como comer, digitar, jogar golf ou cumprimentar alguém com um aperto de mão. O termo gesto, em Wang et al. (2019), diz respeito a movimentos feitos com um ou dois braços, como girar o braço no sentido horário, tocar bateria ou guitarra e bater palmas; em contrapartida, Yang et al. (2020) empregam essa mesma expressão para caracterizar gesticulações com as mãos e dedos; quanto a Canal, Escalera e Angulo (2016), consideram como gestos movimentos feitos com todo o corpo.

Portanto, este trabalho adota o termo "ação" para movimentos feitos pelo corpo humano, considerando da cintura para cima como área de captura e estudo. O grupo de ações, descrito na Seção 2.2, é o meio de comunicação entre humano e robô, para possíveis

aplicações em SAR. Porém, com esse fim, é necessário conhecer os diversos métodos para realizar o reconhecimento e a classificação de ações.

## 1.1 Reconhecimento de Ações

Em razão da sua vasta área de aplicação, reconhecimento de ações vem ganhando mais espaço no campo acadêmico nas últimas décadas. Pode ser usado em espaços públicos para detectar atividades suspeitas, monitoramento médico, avaliação de performance atlética e outras atividades, como interações avançadas entre humano e robô (MOESLUND; HILTON; KRÜGER, 2006) (SEMPENA; MAULIDEVI; ARYAN, 2011) (AGGARWAL; XIA, 2014) (LI et al., 2020).

Convencionalmente, segundo Agahian, Negin e Köse (2019), o reconhecimento de ações humanas pode ser dividido em duas partes:

1. Extração de características com o uso de sensores (Sensor *Radio Frequency Identification*, *InfraRed*, sensor RGB-D e outros);
2. Classificação, utilizando as características extraídas (*Dynamic Time Warping, Hidden Markov Model*, Redes Neurais Artificiais, Espaço Vetoral e outros).

Sensores de reconhecimento de movimento, normalmente, são dispositivos baseados em visão (*Vision Based*) ou em captura de movimento (*Motion Caption*, MoCap), em que o primeiro consiste em utilizar imagens que podem conter cor (RGB) ou informações de profundidade, e, assim, prosseguir para reconhecimento das ações. Já o segundo extrai as informações especificas, como a posição 3D e/ou velocidade das juntas, que podem ser utilizadas através do processamento de informações da imagem ou por marcadores atrelados ao corpo para extrair diretamente suas características (MITRA; ACHARYA, 2007).

Como o reconhecimento de ações baseia-se na implementação prática, ele necessita utilizar diferentes dispositivos de imagem e rastreamento (THOBBI; SHENG, 2010). Atualmente, dispositivos RGB-D de baixo custo, como *Intel RealSense* e *Microsoft Kinect* estão ganhando mais espaço em função do seu custo-beneficio. Esse fato fez com que as pesquisas atuais fossem desenvolvidas utilizando a informação 3D do esqueleto para classificar ações (PATRONA et al., 2018).

Com o esqueleto humano rastreado, devem-se utilizar métodos para criação de um padrão de rótulos. Diversas técnicas são empregadas para tal atividade, como:

1. ***Dynamic Time Warping* (DTW)**, que compara duas séries temporais para encontrar seu nível de similaridade. Essas séries podem ser descritas como um

vetor de características, que representa a orientação e/ou posição das juntas do esqueleto, ou pela imagem de profundidade. Essa técnica também é utilizada para aplicações como mineração de dados e reconhecimento de voz (CELEBI et al., 2013; BARNACHON et al., 2014; RAHEJA et al., 2015; HANG et al., 2017);

2. **Espaço Vetorial**, em que o gesto ou ação pode ser identificado pela aproximação da informação 3D, que é representada com o espaço vetorial sem a necessidade de reconstruir a estrutura 3D. Esse método pode ser aplicado utilizando cada elemento da imagem de profundidade, ou seja, os *pixels* (VEMULAPALLI; ARRATE; CHELLAPPA, 2014; VEMULAPALLI; ARRATE; CHELLAPPA, 2016);

3. ***Hidden Markov Model* (HMM)**, que utiliza a quantificação de uma configuração do sistema através de um número finito de estados discretos, cujos valores armazenados representam a aproximação dinâmica do sistema. Esses estados podem ser o vetor de características, que consiste na posição espacial de cada característica de um objeto ao utilizar um sensor *self-calibrating stereo blob tracker* (ZHANG et al., 2016; KUMAR et al., 2017);

4. **Rede Neural Artificial (RNA)**, é um tipo de treinamento que dependa das entradas com suas respectivas saídas esperadas com diversas estruturas, como Rede Neural Recorrente, bastante utilizada quando as entradas são sequências temporais (NG; RANGANATH, 2002; DU; WANG; WANG, 2015; VEERIAH; ZHUANG; QI, 2015), Rede Neural Convolucional, mais utilizada em situações com processamento de imagens (WANG et al., 2018; YAN; XIONG; LIN, 2018) ou Rede Neural Profunda, que possui diversas camadas e cuja rede requer um banco de dados grande, mas é capaz de extrair as informações de entrada por si mesma (WANG et al., 2015; ORDÓÑEZ; ROGGEN, 2016).

## 1.2 Objetivos e Estrutura

Nesse contexto, o objetivo principal da monografia é desenvolver um algoritmo confiável, usando RNA, a ser empregado em possíveis aplicações de robótica social para tarefas de comunicação com uma máquina.

Com esse intuito, torna-se necessário definir três passos cruciais no desenvolvimento do método. Primeiramente, estudou-se o sensor de movimento utilizado, considerando a disponibilidade e características úteis, como a detecção de juntas de esqueleto humano. Em seguida, foi feita uma análise de base de dados disponível na internet; e, de acordo com o que se pretende neste trabalho, tomou-se a decisão de criar uma base dados específica. Enfim, com a base de dados construída, pode-se aplicar técnicas de pré-processamento para diminuir as dimensões da entrada da RNA, para, assim, executar a rede criada.

A presente monografia está dividida em três capítulos. No Capítulo 2 são descritos o sensor utilizado, a criação da base de dados e as classificações realizadas. No Capítulo 3 são explicitados os resultados obtidos. Por fim, no Capítulo 4, é feita a conclusão.

# 2 Materiais e Métodos

Este capitulo é separado em três seções. Primeiramente, estudou-se o sensor utilizado (*Microsoft Kinect v2.0*); em seguida, foi criada a base de dados e, por fim, esta foi tratada com a intenção de obter o melhor resultado com a técnica proposta (Rede Neural Artificial).

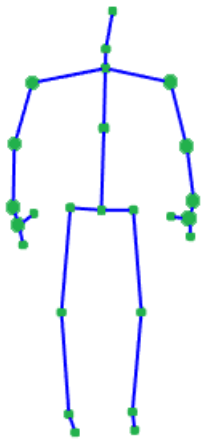
## 2.1 Sensor de Movimento

O sensor *Kinect* foi lançado pela *Microsoft* e, graças a seu baixo custo e sua ampla disponibilidade, tem sido empregado em pesquisas de ciência da computação, eletrônica e engenharia, em atividades que vão desde a ajuda a crianças com autismo até ao auxílio a médicos em cirurgias (ZHANG et al., 2016).

O *Kinect 2.0*, empregado neste estudo, é a versão mais atual do sensor, tendo sido lançado em 2015. Essa versão apresenta diversos sensores, sendo eles uma câmera digital *Red-Green-Blue* (RGB), quatro microfones e uma câmera de profundidade 3D *time-off-flight*, a qual disponibiliza informações espaciais de objetos (X,Y,Z).

As ações humanas podem ser interpretadas por um conjunto de juntas do corpo, sendo estas a junção entre membros – o ombro e o cotovelo – ou extremidades do corpo humano – a cabeça, os pés e os dedos. Porém, não são consideradas fáceis a detecção e a extração de juntas do esqueleto de vídeos RGB normais (YANG; TIAN, 2014). Esse foi um dos pontos principais considerados para escolha do sensor a ser utilizado, já que o *Kinect v2.0* possui uma função integrada de detecção de 25 juntas do corpo humano, como mostra a Figura 2.

Figura 2 – Esqueleto com as 25 juntas destacadas em verde



Fonte: Autor.

A Tabela 1 apresenta algumas especificações desse sensor.

Tabela 1 – Especificação do sensor *Kinect v2.0*

| **Especificação** | **Valor** |
|---|---|
| Câmera RGB | 1920 x 1080 pixels |
| Câmera de profundidade | 512 x 424 pixels |
| Quadros por segundo | 30 FPS |
| Distância de captura | 0.5m – 4.5 m |
| Juntas capturadas | 25 |

Segundo diversas fontes de estudo, como (ASHLEY, 2015; DARBY et al., 2016), para o desenvolvimento de projetos com tal sensor, é recomendada a utilização do *Software Development Kit* (SDK) oficial, disponibilizado pela *Microsoft* para *Windows*. Sua versão mais atual, até a realização deste trabalho, apresenta a possibilidade de conexão com o *Matlab*, *software* que é utilizado para produção desta monografia.

O SDK disponibiliza elementos básicos para tratar de aspectos tais como visão tridimensional, reconhecimento de ações ou gestos, seguimento de esqueleto, áudio, fluxo de dados do sensor, dentre outros. Como o presente trabalho lida com o reconhecimento de ações, esse sensor, aliado ao SDK, é de suma importância para sua realização.

O SDK, juntamente com o *Matlab*, confere ao *Kinect v2.0* a possibilidade de detectar esqueletos, retornando as informações espaciais (X,Y,Z) de 25 juntas do corpo humano, como demonstrado na Figura 2. Por ter precisão suficiente para identificar duas juntas especificas nas mãos (os dedos médio e polegar), o SDK permite utilizar uma função para detectar se as mãos do usuário estão abertas ou fechadas. Tal característica é considerada importante para aplicações em tempo real, como será mostrado na Seção 2.3.

## 2.2 Base de dados

A aplicação e o estudo de técnicas de reconhecimento de gestos e ações utilizam, comumente, base de dados disponíveis na internet. Porém, isso pode gerar empecilhos. Primeiramente, essas bases podem ter uma grande quantidade de dados para processamento, tornando-as inviáveis para algumas aplicações. Em segundo lugar, uma base de dados já disponível carece de flexibilidade para aplicações específicas, ou seja, ela pode não ter ações interessantes para certas finalidades e utilizações.

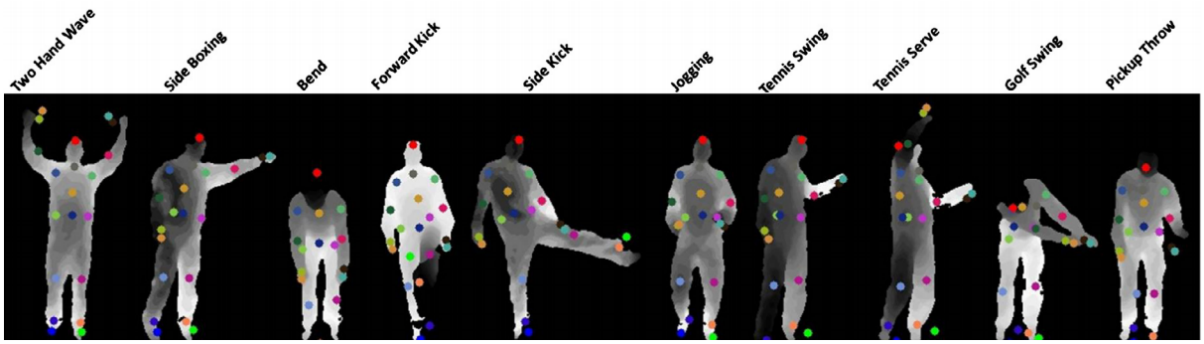
Por esse motivo, essa seção expõe razões para a criação de uma base de dados nova, além de explicitar sua elaboração.

As bases de dados, encontradas na internet, tais como as criadas por Xia, Chen e Aggarwal (2012) e Yang e Tian (2014), têm classes de ações incompatíveis com as tarefas pretendidas para o resultado desta monografia. Alguns exemplos de ações utilizadas nesses

trabalhos podem ser vistos na Figura 3. Pela Figura 3a podemos notar que a base de dados criada por Xia, Chen e Aggarwal (2012) apresenta ações generalizadas, difíceis de serem utilizadas diretamente como comandos para interação humano-robô. Um evento similar pode ser observado na Figura 3b. As ações usadas por Yang e Tian (2014) são voltadas para movimentos ligados a esportes, tais como golf, tênis e futebol. Como o objetivo do trabalho é o desenvolvimento de um algoritmo confiável, usando RNA, a ser aplicado em robótica social para tarefas de comunicação com uma máquina, foi necessário criar uma base de dados própria.

Figura 3 – Exemplos de ações encontradas em base de dados da internet.

(a) Exemplos de ações retirados de (XIA; CHEN; AGGARWAL, 2012). Da esquerda para a direita: andar, ficar em pé, sentar, pegar caixa, carregar caixa.




(b) Exemplos de ações retirados de (YANG; TIAN, 2014). Da esquerda para a direita: tchau com as duas mãos, soco lateral, curvar-se para frente, chute para frente, chute para o lado, caminhada, tacada de tênis, saque de tênis, tacada de golf e arremesso.

A base de dados foi criada utilizando-se o sensor *Kinect v2.0* para extrair a imagem de profundidade e as informações das juntas do esqueleto, o que foi explicitado na Seção 2.1. Nove classes de ações padrões foram escolhidas para integrar a base de dados, o que pode ser visto na Figura 4. Todas as ações partem da posição de descanso quando se está em pé, e são descritas como: um gesto de adeus com a mão direita (Ação A - 4a); levantar a mão direita e desenhar um círculo ao redor da cabeça com a mão (Ação B - 4b); levantar o braço direito até aproximadamente 45 graus com o tronco (Ação C - 4c); um gesto de adeus com a mão esquerda (Ação D - 4d); sobrepor os braços na frente do torso, formando um X (Ação E - 4e); levantar o braço esquerdo até aproximadamente 45 graus com o

tronco (Ação F - 4f); levantar o braço direito como um sinal de "pare" (Ação G - 4g); juntar as mãos na frente da cintura (Ação H - 4h) e levantar o braço para frente com a palma da mão direita para cima, formando um sinal de "vem" (Ação I - 4i).

Além disso, foram adicionadas duas classes à base de dados para serem identificadas como ações/classes neutras, voltadas para a detecção de movimentos naturais do corpo humano. A primeira ação neutra contém amostras do usuário em pé e parado, e a segunda, o usuário andando.

A criação dessas ações neutras é fundamental para empregar o método em aplicações reais. Como a intenção, em robótica social, é que o robô esteja constantemente vendo o usuário, é de suma importância que ele consiga identificar movimentos naturais do corpo humano. Portanto, tais movimentos, quando detectados, são considerados classes neutras que, em possíveis aplicações, não terão comandos atrelados a elas.

Mesmo que o *Kinect* tenha a possibilidade de captura máxima de 30 FPS (quadros por segundo), neste trabalho, ele foi limitado a 15 FPS com a intenção de diminuir o tempo de processamento do algoritmo, além de se abrir a possibilidade de trabalho com sistemas de baixa capacidade de processamento.

Cada ação foi capturada usando uma janela de 1,66 segundos. Esse valor foi escolhido empiricamente, considerado um tempo suficientemente confortável para a realização de todas as ações, principalmente a Ação B, considerada a mais complexa. Toda realização das ações partiu com os usuários parados e em pé, finalizando com a ultima posição referente a cada classe.

Portanto, considerando-se a taxa e a janela de captura, uma amostra de ação possui 25 quadros. Cada quadro consiste em um grupo de características, as juntas do esqueleto, sendo 25 no total, e elas são armazenadas em uma matriz, como a seguir:

$$\mathbf{F}_k = \begin{bmatrix} x_1 & y_1 & z_1 \\ \vdots & \vdots & \vdots \\ x_i & y_i & z_i \end{bmatrix}, \tag{2.1}$$

onde $F_k$ é a matriz de características do $k$-ésimo quadro – de 1 à 25 – e as colunas representam as coordenadas cartesianas 3D $(x_i, y_i, z_i)$ para a $i$-ésima junta do esqueleto – de 1 à 25.

A ação completa é armazenada como a concatenação subsequente da matriz de características de todos os quadros, como a seguir:

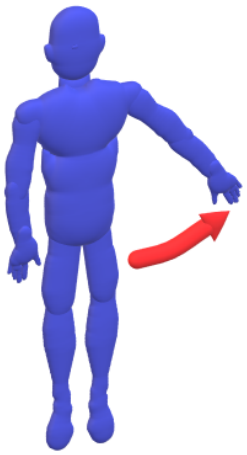
$$\mathbf{A}_{m,n} = [\mathbf{F}_1 \quad \mathbf{F}_2 \ldots \mathbf{F}_k] \tag{2.2}$$

onde $A$ é a matriz de ação composta por características de todos os quadros. Os índices $m$ e $n$ foram usados apenas como etiqueta para representar cada amostra de ação, nos quais

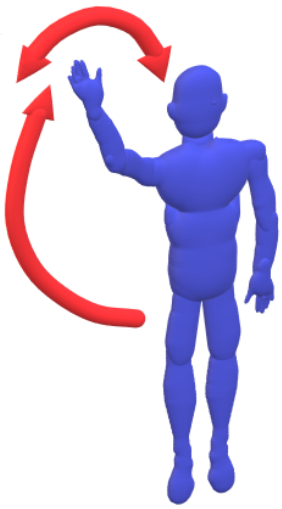
Figura 4 – Ações padrões da base de dados.

(a) Ação A

(b) Ação B

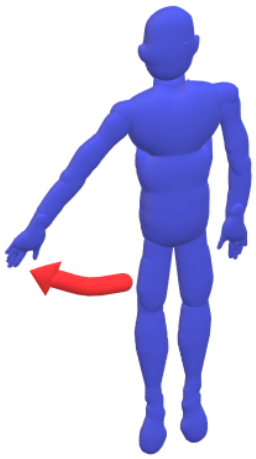


(c) Ação C



(d) Ação D

(e) Ação E



(f) Ação F

(g) Ação G

(h) Ação H

(i) Ação I

$m$ é o rótulo da classe e $n$ é o número de amostra.

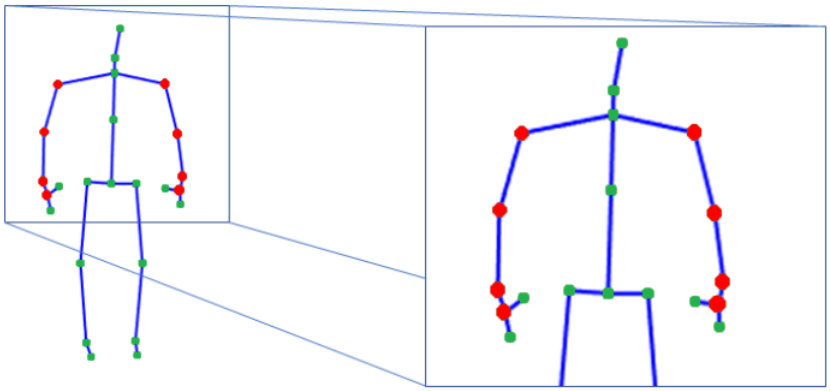
Por fim, a base de dados criada contém um total de 550 amostras de todas as classes descritas (9 classes padrões e 2 classes neutras), sendo 10 amostras de ações coletadas para cada classe por 5 pessoas diferentes, totalizando 50 amostras para cada classe.

## 2.3 Classificação

Antes de trabalhar com a classificação propriamente dita, é necessário realizar um pré-processamento dos dados de entrada (amostras de ações coletadas na base de dados), a fim de classificar diferentes classes de ações com alta taxa precisão e um banco de dados pequeno.

Inicialmente, as dimensões da entrada são reduzidas. As matrizes de ação $A$, Equação (2.2), têm 25 linhas e 75 colunas - sendo as linhas as 25 juntas do esqueleto e as colunas as 25 características com 3 coordenadas cada. Todas as ações são caracterizadas pelo movimento dos braços e das mãos. Então, as juntas consideradas mais importantes, neste caso, são apenas 8, que são referentes ao ombro, cotovelo, pulso e mãos, dos lados esquerdo e direito. Essas juntas estão destacadas em vermelho na Figura 5. Assim, com a intenção de diminuir o número de informações a serem trabalhadas, as matrizes de ação agora têm apenas 8 linhas e 75 colunas.

Figura 5 – Juntas do esqueleto destacadas para aquelas utilizadas na classificação



Fonte: Autor.

Além disso, sabendo-se que uma mesma ação seria igual aos olhos de um observador humano se fosse feita em diferentes pontos de uma sala, quando essas mesmas ações são vistas pelo sensor *Kinect*, elas terão coordenadas totalmente diferentes para suas juntas. Disso resulta uma diferença enorme para duas ações que pertencem à mesma classe, simplesmente por terem sido feitas em lugares distintos. Para evitar esse problema, todas as características são centralizadas em relação às coordenadas da junta do ombro esquerdo,

pois, além de ser considerada uma das juntas importantes para os movimentos, ela possui menor variação na sua posição espacial ao ser comparada com as outras.

O próximo passo é reduzir cada matriz de ação de $8\times75$ para um vetor de autovalores $\lambda$. Em outras palavras, têm-se as componentes principais de $A_{m,n} \cdot A_{m,n}^T \in \mathbb{R}^{8\times8}$, dadas por $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_8$. Assim, essa matriz de ação reduzida, o vetor de autovalores, será a entrada para o treinamento da RNA.

Com os devidos pré-processamentos realizados, pode-se dar inicio ao método de classificação proposto neste trabalho.

### 2.3.1 Classificação *Offline*

A Classificação *Offline*, assim chamada por não ocorrer de forma síncrona com a realização das ações, foi realizada aplicando-se o algoritmo da Rede Neural Artificial à base de dados pré-processada, a fim de se treinar a rede para futura utilização na Classificação *Online*

A entrada da RNA é o seguinte vetor:

$$\textbf{Input} = [\lambda_1 \quad \lambda_2 \quad \dots \quad \lambda_8]^T \tag{2.3}$$

que representa uma amostra de ação.

A Rede Neural utilizada tem apenas uma camada contendo 30 neurônios com todas as camadas conectadas e 11 saídas, que representam, cada uma delas, sua classe específica, a Figura 6 apresenta o esquema da rede criada. A função de ativação utilizada foi a tangente *Sigmoid* para a camada escondida e a função *Softmax* para a camada de classificação. Finalmente, o processo de treinamento utilizado foi a regularização *Bayesiana*, que minimiza uma combinação dos erros quadrados e pesos e, em seguida, determina a combinação correta, de modo a produzir uma rede que generalize bem o reconhecimento de ações.

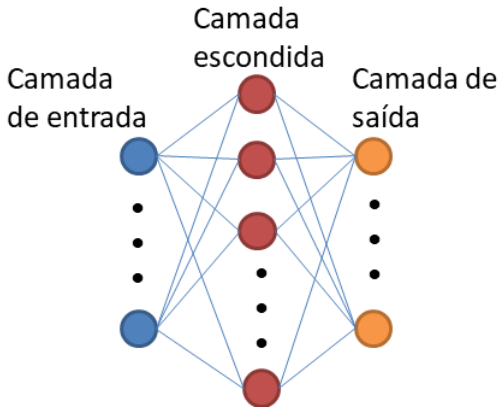
O treinamento da rede foi realizado empregando-se 70% da base de dados como treinamento, 15% como validação e 15% para testes, sendo as amostras escolhidas aleatoriamente dentre todas as presentes na base de dados. Esse procedimento é repetido 10 vezes para se poder ter uma média do desempenho.

A Figura 7 apresenta um fluxograma que sumariza a classificação da RNA proposta.

### 2.3.2 Classificação *Online*

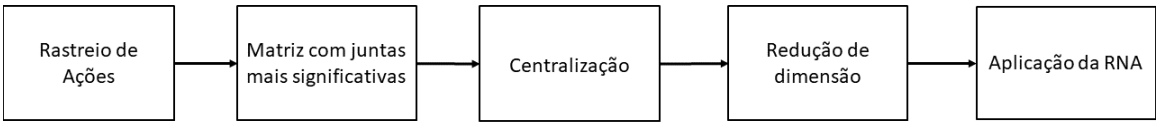
Para realizar os testes de classificação em tempo real, empregou-se a rede treinada para classificação *Offline*. A essa etapa deu-se o nome de Classificação *Online*.

Figura 6 – Representação da RNA utilizada




Fonte: Autor.

Figura 7 – Fluxograma para o algoritmo de reconhecimento de gestos




Fonte: Autor.

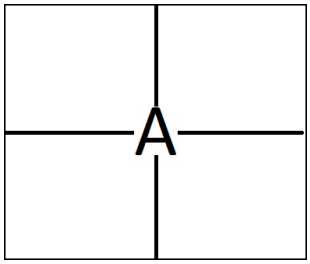
Para evitar situações nas quais o algoritmo possa classificar ações todo o tempo, adicionou-se um gatilho que começa a armazenar quadros antes de classificar. Isso acontece quando o usuário abre uma das mãos, já que o *SDK* do sensor utilizado possui função para detectar se o esqueleto está com as mãos fechadas ou abertas, como é descrito na Seção 2.1.

Para validar o método proposto, foi desenvolvido um sistema para analisar sua performance em tempo real, no qual o sistema solicita que o usuário realize uma ação, e o classificador responde qual ação foi realizada. Apenas com o sensor e o computador, uma imagem sinaliza, de forma aleatória, qual ação o usuário deverá realizar, como mostra a Figura 8. Assim, o computador, de forma simultânea, analisa e armazena se o método acertou ou errou na classificação.

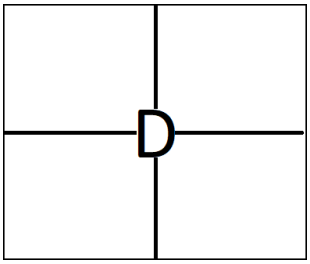
Além desse método, com a intenção de encenar uma aplicação da vida real, criou-se uma simulação com o Jogo da Velha. Para tal, além das ferramentas utilizadas no teste de validação anterior, foi adicionado à estrutura um sistema de navegação usando o robô *Pioneer 3-DX*. Esse sistema foi disponibilizado pelo co-orientador desta monografia, Kevin Braathen de Carvalho.

O Jogo da Velha foi organizado de forma que cada posição a ser preenchida represente uma ação, como mostra a Figura 9. Desse modo, ao realizar uma ação, o robô é direcionado à posição referente. Além disso, no ambiente de aplicação, duas áreas foram

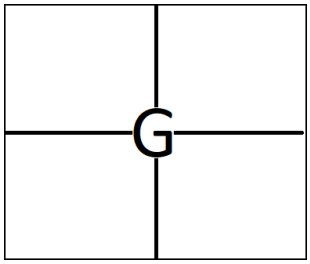
Figura 8 – Imagens para sinalizar qual ação deve ser realizada




(a) Faça a ação A.




(b) Faça a ação D.

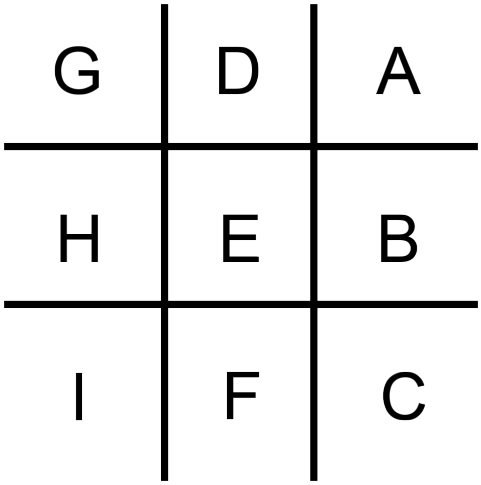



(c) Faça a ação G.

Fonte: Autor.

separadas, uma chamada de Zona de Ação, para realização dos comandos, e outra, Zona de Jogo, onde o robô se move para as posições segundo os comandos.

Figura 9 – Posições referentes às ações para o Jogo da Velha.




Fonte: Autor.

# 3 Resultados e Discussão

Este capítulo apresenta a discussão de resultados das Classificações *Offline* e *Online*, descritos na Seção 2.3 do Capítulo 2. Os resultados são detalhados através da matriz de confusão, em que se pode visualizar a porcentagem das classes de amostra que foram classificadas corretamente. Os valores de cada elemento dessa matriz retratam o número de predições, em porcentagem, das classificações. As linhas representam as classes desejadas, e as colunas, as classes preditas.

Como foi retratado na Seção 2.2 do Capítulo 2, foram levadas em conta as ações neutras, chamadas aqui de Classes J e K, em que a primeira representa uma pessoa em pé e parada, e a segunda, o usuário andando.

A base de dados completa foi criada utilizando 5 pessoas com formas corporais distintas com a intenção de generalizar a classificação e ter uma maior confiabilidade ao lidar com pessoas que não estão na base de dados.

Para as Classificações *Offline*, foram criados três tipos de testes diferentes para analisar como o algoritmo se comporta com diferentes entradas. Para as Classificações *Online*, foram desenvolvidos dois testes para validar o método proposto.

Os nomes dos testes foram escolhidos de acordo com o número de pessoas (esqueletos diferentes) presentes na base de dados de entrada da RNA e com o número de amostras por pessoa. Para os testes *online* foi utilizado como base a rede que apresentou melhor eficiência. Assim, a Tabela 2 apresenta um resumo dos testes realizados.

Tabela 2 – Ensaios realizados para teste e validação do método proposto

| **Testes *Offline*** | | | |
|---|---|---|---|
| **Nome** | **Pessoas** | **Amostras/Pessoa** | **Amostras/Classe** |
| Teste 2P-10A | 2 | 10 | 20 |
| Teste 5P-10A | 5 | 10 | 50 |
| Teste 5P-05A | 5 | 05 | 25 |
| **Testes *Online*** | | | |
| **Nome** | **Rede treinada utilizada** | **Voluntários no teste** | |
| Primeiro teste *online* | 5P-10A | 2 | |
| Segundo teste *online* | 5P-10A | 2 | |

Para examinar como o algoritmo se comporta com pouca variedade de esqueleto para treinamento, o primeiro teste *offline*, chamado **Teste 2P-10A**, foi realizado usando 10 amostras de duas pessoas presentes na base de dados para cada classe, totalizando 20 amostras de rótulo de ação. Para o segundo teste *offline*, chamado **Teste 5P-10A**, foi utilizada a base de dados completa, com 10 amostras de cinco pessoas, presentes na base

de dados, para cada classe, totalizando 50 amostras de cada ação. Assim, pode-se analisar o quão eficiente é o algoritmo no treinamento com um maior número de esqueletos como entrada. Com a intenção de avaliar o algoritmo com uma base de dados pequena, porém, com variedade de esqueleto, para o terceiro teste *offline*, chamado **Teste 5P-05A**, foram empregadas apenas 05 amostras de cinco pessoas diferentes para cada classe, totalizando 25 amostras para cada rótulo de ação. O teste foi realizado com as 05 amostras restantes, não usadas para o treino da RNA, das cinco pessoas.

Para o primeiro teste *online*, foram utilizados dois usuários voluntários diferentes para realizar 30 ações de forma aleatória, como está explicado na Seção 2.3.2. Já no segundo teste *online*, aplicou-se um sistema para simular um Jogo da Velha em tempo real, adicionando um novo sistema ao método, como é mostrado na Seção 2.3.2

É importante destacar que a aplicação deste trabalho tem como proposta o reconhecimento de ações para utilização em aplicações de Robótica Socialmente Assistiva, em que as ações do usuário serão utilizadas como meio de comunicação com o robô. Logo, essa proposta levanta pontos importantes: em primeiro lugar, a taxa de acerto deve ser alta, devido à aplicação dada a cada comando executado. Ações padrões serem confundidas com ações neutras não é considerado ruim, já que, quando é realizado um comando para o agente realizar uma ação e caso ela seja confundida com o usuário em pé e parado ou andando, o robô entenderá tal ação como neutra, e o usuário terá de repeti-la. Em segundo lugar, as ações neutras devem ter maior precisão ao serem comparadas com as padrões, já que se o usuário está em pé e parado ou andando, e caso ela seja erroneamente classificada como uma ação relacionada com um comando, o robô alterará sua tarefa, tornando, assim, a aplicação inviável.

## 3.1 Classificação *Offline*

### 3.1.1 Teste 2P-10A

Para obter o resultado exposto na Tabela 3, foram realizados 10 testes de validação e, depois, tirada sua média para garantir a confiabilidade do resultado. Para tal teste, foi aplicada uma base de dados de entrada na rede com um total de 220 amostras, sendo 10 de cada classe de ação, coletadas por duas pessoas diferentes.

Através da análise da Tabela 3 a Classe K se mostrou mais problemática, pois, ao realizar a Ação F e G, 4% das vezes, elas foram classificadas como K. Por ser uma classe neutra, esse não é um grande problema, pois, conforme a intenção das possíveis aplicações, ações neutras não representam um comando, o que tornará necessário repetir o comando.

Além disso, podemos notar que o método, junto com o teste proposto, possui alta taxa de acerto, de 98,3%, usando apenas duas pessoas na base de dados, sendo considerado

Tabela 3 – Matriz de confusão para o teste 2P-10A: 98,3% taxa de acerto

| Classes Preditas \ Classes Desejadas | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 100 | - | - | - | - | - | - | - | - | 2 | - |
| **B** | - | 100 | - | - | - | - | - | - | - | - | - |
| **C** | - | - | 100 | - | - | - | - | - | - | - | - |
| **D** | - | - | - | 100 | - | - | - | - | - | - | - |
| **E** | - | - | - | - | 100 | - | - | - | - | - | 1 |
| **F** | - | - | - | - | - | 94 | - | - | 2 | - | 2 |
| **G** | - | - | - | - | - | - | 96 | - | - | - | - |
| **H** | - | - | - | - | - | - | - | 100 | - | - | - |
| **I** | - | - | - | - | - | 2 | - | - | 98 | - | 2 |
| **J** | - | - | - | - | - | - | - | - | - | 98 | - |
| **K** | - | - | - | - | - | 4 | 4 | - | - | - | 95 |

esse valor um representativo de boa performance.

### 3.1.2 Teste 5P-10A

De forma similar ao primeiro teste, foram realizados 10 testes de validação, dos quais a média é apresentada na Tabela 4. Aqui, usou-se um total de 550 amostras, sendo cada 10 amostras de uma classe de ação, coletadas por 5 pessoas diferentes.

Tabela 4 – Matriz de confusão para o teste 5P-10A: 99,4% taxa de acerto

| Classes Preditas \ Classes Desejadas | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 98,6 | 0,2 | - | - | - | - | 0,2 | - | - | - | - |
| **B** | 0,4 | 99,8 | - | - | - | - | - | - | - | - | - |
| **C** | - | - | 100 | - | - | - | - | - | - | - | 1,2 |
| **D** | - | - | - | 99,2 | - | - | - | - | - | - | - |
| **E** | - | - | - | - | 100 | - | - | - | - | - | - |
| **F** | 0,5 | - | - | - | - | 99 | - | - | - | - | 0,2 |
| **G** | 0,5 | - | - | 0,4 | - | - | 99,6 | - | 0,2 | - | - |
| **H** | - | - | - | - | - | - | - | 99,8 | - | - | - |
| **I** | - | - | - | - | - | - | 0,2 | - | 99,8 | - | - |
| **J** | - | - | - | - | - | - | - | - | - | 100 | - |
| **K** | - | - | - | 0,4 | - | 1 | - | 0,2 | - | - | 98,6 |

Analisando a Tabela 4 juntamente com a Tabela 3, é possível notar que adicionar pessoas com composições corporais diferentes não alterou de forma significativa a capacidade de aprendizagem e classificação do algoritmo. Além disso, as classes F e G, antes consideradas problemáticas, agora tiveram uma maior taxa de acerto, o que deve ter ocorrido devido à maior variação de esqueletos na base de dados de entrada.

Portanto, obteve-se uma taxa de acerto de 99,4% nesse segundo teste. Isso indica que, ao aumentar o número de usuários do sistema, não houve queda no desempenho do classificador. Logo, o sistema proposto possui capacidade de generalização.

### 3.1.3 Teste 5P-05A

Assim como nos anteriores, para o terceiro teste, foi feita uma validação cruzada, aplicando 10 vezes o algorítimo proposto e, em seguida, foi tirada a média dos resultados. O resultado obtido desse teste está na Tabela 5. Aqui, foi utilizado um total de 225 amostras, sendo 25 de cada classe de ação, coletadas por cinco pessoas diferentes, lembrando-se que o teste de performance foi aplicado às outras 5 amostras não utilizadas para o treinamento dessa RNA.

Tabela 5 – Matriz de confusão para o teste 5P-05A: 97,5% taxa de acerto

| | | Classes Desejadas | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** |
| Classes Preditas | **A** | 97 | - | - | - | - | - | - | - | - | - | - |
| | **B** | 3 | 100 | - | 2,4 | - | - | - | - | 0,3 | - | - |
| | **C** | - | - | 96,3 | - | - | - | - | - | - | - | - |
| | **D** | - | - | - | 96,6 | 0,3 | - | 3 | - | 0,7 | - | - |
| | **E** | - | - | - | - | 99 | - | - | - | - | - | - |
| | **F** | - | - | - | - | - | 96,3 | - | - | - | - | - |
| | **G** | - | - | - | - | - | - | 96,6 | - | 0,4 | - | - |
| | **H** | - | - | - | - | - | - | - | 96,6 | - | - | 0,3 |
| | **I** | - | - | - | 1 | 0,7 | - | 0,4 | - | 95,3 | - | - |
| | **J** | - | - | 0,3 | - | - | - | - | - | 2,3 | 100 | 0,7 |
| | **K** | - | - | 3,4 | - | - | 3,7 | - | 3,4 | 1 | - | 99 |

Pela Tabela 5, é possível notar uma menor precisão geral quando comparada com as Tabelas 3 e 4. Porém, é importante salientar que a maioria dos erros de classificação ocorreu quando uma ação padrão foi confundida com uma ação neutra. Apenas 1,1% das amostras foi confundida com uma ação padrão diferente, ou seja, apenas 1,1% dos comandos faria algo diferente do que foi designado a ele, em alguma possível aplicação. É importante destacar que os resultados desta tabela são de amostras que a RNA não utilizou para treinamento, ou seja, as ações classificadas eram desconhecidas para a RNA.

Por fim, vale dizer que o método apresenta uma boa performance mesmo com uma base de dados pequena, já que se obteve uma taxa de acerto de 97,5%.

## 3.2 Classificação *Online*

### 3.2.1 Primeiro Teste

A RNA utilizada para esse teste foi criada usando todas as 50 amostras, sendo 10 de cada pessoa para cada classe, já que, segundo os resultados anteriores, para tal entrada pode-se obter uma maior taxa de acerto. Os testes foram realizados por duas pessoas voluntárias que estavam na base de dados, e cada uma realizou cada ação 30 vezes em ordem aleatória para evitar que a mesma ação fosse solicitada múltiplas vezes em sequência. Os resultados desse experimento são apresentados na Tabela 6.

Tabela 6 – Matriz de confusão para o primeiro teste *online*: 98% taxa de acerto

| | | Classes Desejadas | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** |
| Classes Preditas | **A** | 100 | 5 | - | - | - | - | 6,7 | - | - | - | - |
| | **B** | - | 91,6 | - | - | - | - | - | - | - | - | - |
| | **C** | - | - | 96,7 | - | - | - | - | - | - | - | 1,7 |
| | **D** | - | - | - | 100 | - | - | - | - | - | - | - |
| | **E** | - | - | - | - | 98,3 | - | - | - | - | - | - |
| | **F** | - | - | - | - | - | 100 | - | - | - | - | - |
| | **G** | - | - | - | - | - | - | 93,3 | - | - | - | - |
| | **H** | - | - | - | - | 1,7 | - | - | 100 | - | - | - |
| | **I** | - | 1,6 | - | - | - | - | - | - | 100 | - | - |
| | **J** | - | - | - | - | - | - | - | - | - | 100 | - |
| | **K** | - | 1,8 | 3,3 | - | - | - | - | - | - | - | 98,3 |

Toda vez que alguém realiza a ação, ela parecerá diferente e se diferenciará ainda mais das amostras de treinamento, já que a ação-gatilho agora é necessária. Em outras palavras, durante os testes *online*, o usuário deve abrir suas mãos para iniciar a captura de ações. Essas ações são um pouco diferentes quando comparadas às ações capturadas para o banco de dados, em que não é necessário utilizar o gatilho manual. A taxa de acerto foi alta com algumas ações padrões sendo confundidas com ações neutras, como mostra a Tabela 6.

Comparando a Tabela 6 com a Tabela 4, é possível aproximar os resultados, o que sugere que a diferença entre uma ação na qual se usa ou não o gatilho manual não é significante para a performance do método proposto.

Desse modo, verificou-se que, a partir do primeiro teste *online*, o método apresenta uma boa performance e capacidade de generalização, representadas pela taxa de acerto de 98%.

### 3.2.2 Segundo Teste

Com o intuito de simular uma possível aplicação simples para o método proposto, foi empregado um algoritmo para reproduzir um Jogo da Velha. Para tal, foi adicionado um sistema com a utilização do robô *Pioneer 3-DX*.

Três partidas foram jogadas entre duas pessoas; nove classes de ações padrões representam as nove posições disponíveis para o jogo. Para representar o final da partida e o robô voltar à posição inicial, o vencedor teve de realizar a mesma ação duas vezes seguidas.

A chamada Zona de Ação representa o local onde o usuário realiza as ações, e a Zona de Jogo é o espaço para o robô se mover, de forma que o jogo possa acontecer. As classes de ações para cada posição, bem como a estrutura criada para o jogo, estão apresentadas na Figura 10. Uma demonstração do teste pode ser vista em: https://youtu.be/5p2SaQhXvks.

Figura 10 – Ambiente criado para simulação de aplicação.


Fonte: Autor.

O vídeo, sem cortes, com três partidas de Jogo da Velha, mostra um uso do algoritmo de classificação, com duas pessoas diferentes. Buscando-se uma melhor experiência do

espectador, é feita uma descrição para mostrar qual partida está começando, e diversas partes foram aceleradas.

Pelo vídeo, pode-se notar que nenhuma ação foi confundida, mesmo quando duas ações foram realizadas em sequência para sinalizar o início de uma nova partida. Isso mostra a alta confiabilidade do algoritmo.

# 4 Considerações Finais

Nesta monografia foi proposto um método de fácil implementação para classificação de ações utilizando Redes Neurais Artificiais, baseado nas juntas de um esqueleto humano, coletado pelo sensor *Kinect v2.0*, para ser empregado em possíveis aplicações de Robótica Socialmente Assistiva. O método depende da redução de dimensão das amostras da base de dados criada para, assim, serem aplicadas à RNA e serem aprendidos os padrões de classificação predefinidos, nos quais, com apenas 5 a 10 amostras de cada classe para cada pessoa, obtêm-se de 97,5% a 99,4% de taxa de acerto.

Os resultados mostram alta média de acerto do treinamento até mesmo para ações neutras. Além disso, o algoritmo mostra que, mesmo com cinco esqueletos diferentes na base de dados, sua performance não é diminuída se for comparada quando se usam apenas dois esqueletos diferentes, fazendo com que o método proposto seja confiável e que haja a possibilidade de usar um banco de dados pequeno para ser efetivo. Portanto, o desenvolvimento do algoritmo pode levar a uma flexível e fácil implementação de uma interface humano-robô, utilizando o reconhecimento de ações, em razão da sua simplicidade e do fato de não necessitar uma base de dados externa para seu treinamento. Ademais, é notável como o método permite incorporar novos usuários com facilidade, uma vez que demanda um baixo número de repetições, para que as amostras possam ser adicionadas à base de dados e um novo usuário possa ser reconhecido.

Para futuros trabalhos pode-se estudar a variação de características do método e das simulações. Como exemplo, variar o valor do FPS para taxas maiores e menores que 15 na captura das ações; variar e aplicar métodos à RNA utilizada para encontrar um valor ótimo para o número de neurônios.

Além disso, ao falar dos testes aplicados, pode-se avaliar o comportamento dos testes *online* ao utilizar pessoas que não estão presentes na base de dados. Pode-se trabalhar diretamente com o comitê de ética para que seja instalado o sistema em algum ambiente público e, assim, poderá ser solicitado às pessoas que façam alguma ação, podendo ser utilizado para ampliar a base de dados ou apenas para validação.

Em relação às possíveis aplicações reais do método, nos próximos trabalhos pode-se tratar com elas, em cenários nos quais a SAR é aplicável. Como exemplo, pode-se lidar com casos em que ações monótonas ou perigosas para seres humanos possam ser realizadas por máquinas controladas usando ações; trabalhar com possíveis aplicações com a intenção de evitar o contato com superfícies prováveis de estarem contaminadas; ser útil em hospitais para transportar dejetos, materiais ou até mesmo remédios. Para mais, existe a possibilidade de adicionar usuários ao sistema, fazendo uma capacitação através

da captura de amostras das classes.

# Referências

AGAHIAN, S.; NEGIN, F.; KÖSE, C. An efficient human action recognition framework with pose-based spatiotemporal features. *Engineering Science and Technology, an International Journal*, Elsevier, 2019. Citado na página 15.

AGGARWAL, J. K.; XIA, L. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, Elsevier, v. 48, p. 70–80, 2014. Citado na página 15.

ASHLEY, J. Beginning kinect programming: with the kinect for windows v2 sdk. Apress, 2015. Citado na página 19.

BARNACHON, M. et al. Ongoing human action recognition with motion capture. *Pattern Recognition*, Elsevier, v. 47, n. 1, p. 238–247, 2014. Citado na página 16.

BASILIO, V. T.; CARVALHO, K. B. d.; BRANDãO, A. S. Reconhecimento de ações por rna em aplicações de robótica social. In: GALOÁ. *Anais do 14° Simpósio Brasileiro de Automação Intelingente*. Ouro Preto, MG, Brasil, 2019. p. 23–28. Citado na página 13.

BENITTI, F. B. V. Exploring the educational potential of robotics in schools: A systematic review. *Computers & Education*, Elsevier, v. 58, n. 3, p. 978–988, 2012. Citado na página 13.

CANAL, G.; ESCALERA, S.; ANGULO, C. A real-time human-robot interaction system based on gestures for assistive scenarios. *Computer Vision and Image Understanding*, Elsevier, v. 149, p. 65–77, 2016. Citado na página 14.

CELEBI, S. et al. Gesture recognition using skeleton data with weighted dynamic time warping. In: VISAPP. *VISAPP (1)*. Barcelona, Spain, 2013. p. 620–625. Citado na página 16.

DARBY, J. et al. An evaluation of 3d head pose estimation using the microsoft kinect v2. *Gait & posture*, Elsevier, v. 48, p. 83–88, 2016. Citado na página 19.

DU, Y.; WANG, W.; WANG, L. Hierarchical recurrent neural network for skeleton based action recognition. In: IEEE. *Proceedings of the IEEE conference on computer vision and pattern recognition*. Boston, MA, USA, 2015. p. 1110–1118. Citado na página 16.

ERIKSSON, J. Hands-off robotics for post-stroke arm rehabilitation. *Technical Report*, 2004. Citado na página 13.

FEIL-SEIFER, D.; MATARIC, M. J. Defining socially assistive robotics. In: IEEE. *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* Chicago, IL, USA, 2005. p. 465–468. Citado na página 13.

FONG, T.; NOURBAKHSH, I.; DAUTENHAHN, K. A survey of socially interactive robots. *Robotics and autonomous systems*, Elsevier, v. 42, n. 3-4, p. 143–166, 2003. Citado na página 13.

HANG, C. et al. Dynamic gesture recognition method based on improved dtw algorithm. In: IEEE. *2017 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)*. Wuhan, China, 2017. p. 71–74. Citado na página 16.

HEIDARI, N.; IOSIFIDIS, A. Temporal attention-augmented graph convolutional network for efficient skeleton-based human action recognition. *arXiv preprint arXiv:2010.12221*, 2020. Citado na página 14.

HUTTENRAUCH, H.; EKLUNDH, K. S. Fetch-and-carry with cero: Observations from a long-term user study with a service robot. In: IEEE. *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*. Berlin, Germany, 2002. p. 158–163. Citado na página 13.

KAHN, L. E. et al. Comparison of robot-assisted reaching to free reaching in promoting recovery from chronic stroke. In: IOS PRESS. *Proceedings of the international conference on rehabilitation robotics*. Istanbul, Turkey, Turkey, 2001. p. 39–44. Citado na página 13.

KANDA, T. et al. Development and evaluation of interactive humanoid robots. *Proceedings of the IEEE*, IEEE, v. 92, n. 11, p. 1839–1850, 2004. Citado na página 13.

KUMAR, P. et al. Coupled hmm-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*, Elsevier, v. 86, p. 1–8, 2017. Citado na página 16.

LI, J. et al. Spatio-temporal deformable 3d convnets with attention for action recognition. *Pattern Recognition*, Elsevier, v. 98, p. 107037, 2020. Citado 2 vezes nas páginas 14 e 15.

MITRA, S.; ACHARYA, T. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, IEEE, v. 37, n. 3, p. 311–324, 2007. Citado na página 15.

MOESLUND, T. B.; HILTON, A.; KRÜGER, V. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, Elsevier, v. 104, n. 2-3, p. 90–126, 2006. Citado na página 15.

MONTEMERLO, M. et al. Experiences with a mobile robotic guide for the elderly. *AAAI/IAAI*, v. 2002, p. 587–592, 2002. Citado na página 13.

NG, C. W.; RANGANATH, S. Real-time gesture recognition system and application. *Image and Vision computing*, Elsevier, v. 20, n. 13-14, p. 993–1007, 2002. Citado na página 16.

ORDÓÑEZ, F.; ROGGEN, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, Multidisciplinary Digital Publishing Institute, v. 16, n. 1, p. 115, 2016. Citado na página 16.

PASCARELLA, G. et al. Covid-19 diagnosis and management: a comprehensive review. *Journal of Internal Medicine*, Wiley Online Library, 2020. Citado na página 14.

PATRONA, F. et al. Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognition*, Elsevier, v. 76, p. 612–622, 2018. Citado na página 15.

RAHEJA, J. et al. Robust gesture recognition using kinect: A comparison between dtw and hmm. *Optik*, Elsevier, v. 126, n. 11-12, p. 1098–1104, 2015. Citado na página 16.

SEMPENA, S.; MAULIDEVI, N. U.; ARYAN, P. R. Human action recognition using dynamic time warping. In: IEEE. *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*. Bandung, Indonesia, 2011. p. 1–5. Citado na página 15.

TAPUS, A.; MAJA, M.; SCASSELLATTI, B. The grand challenges in socially assistive robotics. *IEEE Robotics and Automation Magazine*, v. 14, n. 1, p. N–A, 2007. Citado na página 13.

THOBBI, A.; SHENG, W. Imitation learning of hand gestures and its evaluation for humanoid robots. In: IEEE. *The 2010 IEEE International Conference on Information and Automation*. Harbin, China, 2010. p. 60–65. Citado na página 15.

VEERIAH, V.; ZHUANG, N.; QI, G.-J. Differential recurrent neural networks for action recognition. In: IEEE. *Proceedings of the IEEE international conference on computer vision*. Santiago, Chile, 2015. p. 4041–4049. Citado na página 16.

VEMULAPALLI, R.; ARRATE, F.; CHELLAPPA, R. Human action recognition by representing 3d skeletons as points in a lie group. In: IEEE. *Proceedings of the IEEE conference on computer vision and pattern recognition*. Washington, DC, USA, 2014. p. 588–595. Citado na página 16.

VEMULAPALLI, R.; ARRATE, F.; CHELLAPPA, R. R3dg features: Relative 3d geometry-based skeletal representations for human action recognition. *Computer Vision and Image Understanding*, Elsevier, v. 152, p. 155–166, 2016. Citado na página 16.

WANG, P. et al. Action recognition from depth maps using deep convolutional neural networks. *IEEE transactions on human-machine systems*, IEEE, v. 46, n. 4, p. 498–509, 2015. Citado na página 16.

WANG, P. et al. Depth pooling based large-scale 3-d action recognition with convolutional neural networks. *IEEE Transactions on Multimedia*, IEEE, v. 20, n. 5, p. 1051–1061, 2018. Citado na página 16.

WANG, Q. et al. Space-time event clouds for gesture recognition: from rgb cameras to event cameras. In: IEEE. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Aspen, CO, 2019. p. 1826–1835. Citado na página 14.

XIA, L.; CHEN, C.-C.; AGGARWAL, J. K. View invariant human action recognition using histograms of 3d joints. In: IEEE. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Providence, RI, 2012. p. 20–27. Citado 2 vezes nas páginas 19 e 20.

YAN, S.; XIONG, Y.; LIN, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI. *Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, Louisiana, USA, 2018. Citado na página 16.

YANG, S. et al. Collaborative learning of gesture recognition and 3d hand pose estimation with multi-order feature analysis. In: ECCV. *Proceedings of the European Conference on Computer Vision (ECCV)*. Glasgow, UK, 2020. Citado na página 14.

YANG, X.; TIAN, Y. Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, Elsevier, v. 25, n. 1, p. 2–11, 2014. Citado 3 vezes nas páginas 18, 19 e 20.

ZHANG, X.-H. et al. Improvement of dynamic hand gesture recognition based on hmm algorithm. In: IEEE. *2016 International Conference on Information System and Artificial Intelligence (ISAI)*. Hong Kong, China, 2016. p. 401–406. Citado 2 vezes nas páginas 16 e 18.