

Drone Controlado por Gestos Manuais Classificados por Rede Neural

Henrique Antônio Guanais Corneau

Grad. em Eng. Mecatrônica

Univ. Fed. do Rio Grande do Norte

Natal, Brasil

<https://orcid.org/0009-0004-8815-7842>

João Victor do N. R. de Andrade

Grad. em Eng. de Computação

Univ. Fed. do Rio Grande do Norte

Natal, Brasil

<https://orcid.org/0009-0001-5004-578X>

Ramon Vinícius Ferreira de Souza

Grad. em Eng. de Computação

Univ. Fed. do Rio Grande do Norte

Natal, Brasil

<https://orcid.org/0009-0000-5076-1842>

Bruno Marques Ferreira da Silva

Prog. de Pós-Grad. em Eng. Mecatrônica

Univ. Fed. do Rio Grande do Norte

Natal, Brasil

<https://orcid.org/0000-0002-7780-7254>

Abstract—Este trabalho apresenta um sistema de controle para drones baseado em gestos manuais, utilizando redes neurais para a detecção e classificação dos gestos. A solução proposta foi implementada no drone DJI Tello, um modelo comercial de baixo custo e programação aberta. O sistema desenvolvido permite a execução de quatro ações pré-determinadas: decolagem, pouso, permanência em posição fixa e rastreamento da mão. Os testes foram conduzidos em um ambiente controlado, sem interferências externas e com iluminação adequada. Os resultados obtidos demonstraram a eficácia do sistema, validando o uso de redes neurais para o controle intuitivo de drones acessíveis e de fácil implementação.

Index Terms—Drones; Gestos Manuais; Redes Neurais.

I. INTRODUÇÃO

As redes neurais são modelos computacionais inspirados no funcionamento do cérebro humano, na medida em que é capaz de aprender padrões e realizar tarefas complexas ao serem treinadas, como classificação de imagens, reconhecimento de gestos e controle de sistemas autônomos [1].

No campo do reconhecimento de pontos de referência do corpo humano em imagens e vídeos, o MediaPipe [2] se destaca como uma ferramenta notável. Desenvolvido pelo Google, este software de código aberto simplifica a construção de fluxos de processamento de dados para identificar e acompanhar pontos de referência em tempo real. Esses pontos, que denominam-se *landmarks*, são pontos específicos em um objeto, como a mão ou o rosto e seus respectivos elementos, como olhos, boca, nariz, palma e dedos [3]. O MediaPipe é particularmente útil para analisar imagens e vídeos com precisão, pois permite armazenar apenas os *landmarks*, em vez de imagens completas, aumentando significativamente a eficiência, como embasado em [4]. No presente trabalho, esses *landmarks* foram utilizados como base para a criação de um conjunto de dados com 4.927 amostras, coletadas manualmente por um usuário e distribuídas entre quatro categorias de gestos. Cada amostra é formada por 21 desses pontos de referência, cada um contendo as coordenadas bidimensionais

(x, y) normalizadas, correspondentes à posição dos pontos característicos da mão em cada imagem capturada.

As tecnologias de reconhecimento gestual já são uma realidade, como exemplificado pelo Kinect da Microsoft, que demonstra a versatilidade desses recursos. Tal fato abre caminho para aplicações em diversos setores, incluindo o monitoramento ambiental com câmeras e drones, permitindo a interpretação de gestos e expressões para identificar possíveis situações de risco [5].

Apesar dos avanços tecnológicos, ainda há desafios significativos no controle de drones por gestos. Muitas abordagens exigem dispositivos adicionais ou demandam alto poder computacional, o que limita sua aplicabilidade em sistemas embarcados e em contextos com recursos limitados. Diante disso, este trabalho propõe-se a enfrentar a seguinte problemática: seria possível desenvolver um sistema de controle gestual para drones que seja eficiente e de baixo custo computacional, utilizando tecnologias acessíveis?

Neste trabalho, foi desenvolvido um sistema de controle para drones baseado no reconhecimento de gestos, integrando a detecção da mão por meio do MediaPipe com uma rede neural do tipo Multi-Layer Perceptron (MLP), treinada usando o TensorFlow para classificar diferentes comandos (gestos). Com base na interpretação desses comandos, o drone é capaz de executar ações como decolagem, pouso e rastreamento da mão. O objetivo principal deste estudo foi explorar a viabilidade do uso de redes neurais para o controle de veículos aéreos não tripulados, avaliando sua precisão e desempenho.

II. TRABALHOS RELACIONADOS

Já existem na bibliografia diversos trabalhos que têm como objetivo propor diferentes formas de controlar um drone, eliminando métodos convencionais, como o rádio controle. Referências como [6] e [7] sugerem o uso das mãos para esse controle. No entanto, esses estudos fazem uso de dispositivos externos, que são, respectivamente, um Leap Motion

Controller e um Wearable Device equipado com IMU, o que pode ser útil em alguns cenários. Entretanto, a necessidade desses dispositivos adiciona uma etapa extra na fabricação ou aquisição do sistema.

Outros trabalhos, como [8], utilizam Convolutional Neural Networks (CNNs) para a detecção de gestos de mão, treinados a partir de imagens. O trabalho em específico utilizou SSD (Single Shot MultiBox Detector), que utiliza CNNs como base para realizar detecção de objetos em imagens [9].

O trabalho de [10] apresenta uma abordagem eficiente para o reconhecimento de gestos dinâmicos de mão baseada em redes neurais convolucionais tridimensionais (3D CNNs). O classificador proposto realiza a fusão de volumes de movimento, combinando mapas de profundidade normalizados e valores de gradiente de imagem extraídos de sequências temporais de vídeo.

Essas abordagens podem ser vantajosas em cenários onde é necessário extrair informações detalhadas da mão detectada. No entanto, pode demandar um alto custo computacional.

Já [3] adotou uma abordagem baseada no cascade classifier de Haar features, na qual primeiro foram extraídas as Haar features das imagens de gestos de mão e, em seguida, o algoritmo AdaBoost foi utilizado para o treinamento do classificador. Essa metodologia se destaca pela eficácia em cenários com imagens de baixa qualidade, como em condições de pouca iluminação ou baixa resolução. Além disso, apresenta menor complexidade e custo computacional em comparação com redes neurais convolucionais (CNNs), sendo, portanto, ideal para aplicações em tempo real e ambientes com recursos limitados.

Dessa forma, este trabalho propõe uma abordagem diferente para a detecção de gestos de mão. Em vez de treinar uma rede neural para reconhecer gestos diretamente, utiliza-se uma rede já existente para detectar a mão e seus *landmarks*, que são pontos específicos nas articulações e regiões chave da mão, como as pontas dos dedos e a palma. Em seguida, uma nova rede neural é treinada para classificar os gestos com base nas coordenadas desses *landmarks*.

Essa abordagem é mais eficiente em comparação com as propostas dos trabalhos mencionados anteriormente, se destacando pela facilidade de implementação e pelo baixo custo computacional.

III. MÉTODOS

A abordagem proposta para o controle de drones por gestos de mão envolve três etapas principais: recebimento e processamento da imagem, reconhecimento dos gestos e envio de comandos ao drone. Para isso, o método combina técnicas de visão computacional e aprendizado de máquina, utilizando o OpenCV para o processamento de imagens, MediaPipe para a detecção da mão e extração de pontos-chave, os *landmarks*, e TensorFlow para a classificação dos gestos. O desenvolvimento de todo o projeto foi realizado na linguagem Python e o drone utilizado neste trabalho é o DJI Tello, um modelo compacto e acessível amplamente empregado em aplicações educacionais e de pesquisa.

A captura da imagem é realizada pela câmera do drone, sendo transmitida para um computador via Wi-Fi. O fluxo de vídeo recebido é pré-processado, e em seguida utiliza-se o MediaPipe Hands, identificando a posição da mão na imagem e fornecendo um conjunto de 21 pontos-chave representando articulações e as extremidades dos dedos. Esses pontos são organizados e formatados em formato vetorial, servindo de entrada para o modelo de aprendizado de máquina.

O reconhecimento dos gestos é realizado por meio de um modelo de classificação baseado em TensorFlow, utilizando a implementação disponível no repositório do GitHub *hand-gesture-recognition-mediapipe* [11]. O modelo foi treinado com um conjunto de dados contendo diferentes posturas da mão, cada uma associada a um comando específico. A entrada do modelo consiste nas coordenadas dos pontos-chave extraídos pelo MediaPipe, que são processados e classificados em uma das categorias definidas, como *tracking*, *stop*, *takeoff* ou *land*. A inferência ocorre em tempo real, garantindo uma resposta rápida aos comandos do usuário.

Após a classificação do gesto, o sistema converte o resultado em um comando correspondente para o DJI Tello. A comunicação com o drone é feita via protocolo UDP, utilizando a Tello SDK, que permite o envio de strings formatadas para a execução de ações específicas. Cada gesto reconhecido é mapeado para um comando distinto, sendo eles: palma aberta para *tracking* (seguindo a abordagem implementada em [12]), mão fechada para *stop*, gesto de "V" com o indicador e o médio para decolar e gesto de "L" com o polegar e o indicador para pousar o drone. Os gestos possíveis e seus respectivos comandos são mostrados na Fig. 1. Para evitar comandos indesejados devido a detecções errôneas, é aplicada uma lógica de filtragem que considera apenas gestos mantidos por um período mínimo de tempo antes da execução do comando. Um fluxograma que explica a abordagem escolhida é mostrado na Fig. 2.

IV. RESULTADOS EXPERIMENTAIS

Para avaliar o desempenho do sistema de controle por gestos, foram conduzidos experimentos em um ambiente controlado, garantindo condições adequadas para a captura e processamento das imagens. O drone DJI Tello foi utilizado nos testes, operando em um ambiente interno com boa iluminação, assegurando a precisão na detecção da mão e no reconhecimento dos gestos. Um vídeo que mostra a execução do controle pode ser observado em [13].

O sistema foi executado em um computador equipado com um processador Intel® Core™ i5-1135G7, 32 GB de memória RAM e sem GPU dedicada. O ambiente de desenvolvimento utilizou Python 3.10, com as bibliotecas MediaPipe, TensorFlow e OpenCV. O modelo de reconhecimento de gestos foi previamente treinado e integrado ao pipeline de inferência, possibilitando a execução dos comandos em tempo real.

O modelo foi treinado com um conjunto de dados contendo 4 classes, totalizando 4.927 amostras, distribuídas de forma equilibrada entre as categorias. O dataset foi dividido em um

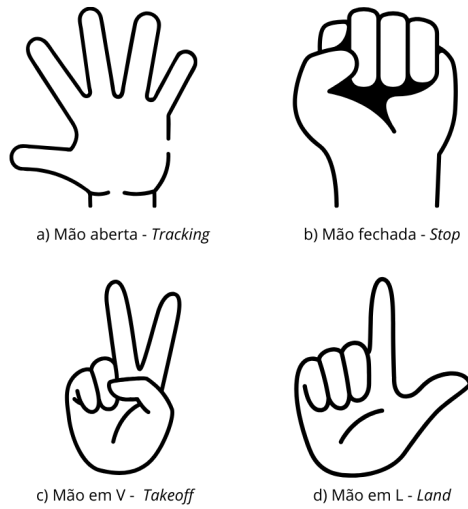


Fig. 1. Gestos e comandos.

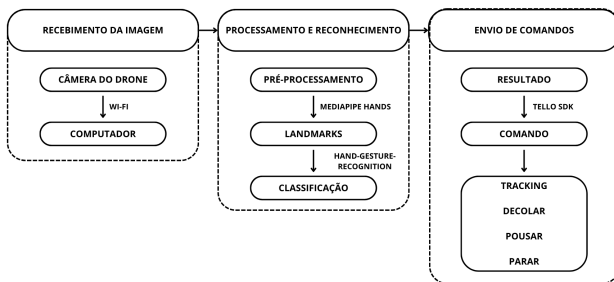


Fig. 2. Fluxograma do processo.

conjunto de treino, teste e validação, contendo, respectivamente, 75%, 12.5% e 12.5% dos dados. O treinamento foi realizado ao longo de 103 épocas, com um batch size de 128, buscando otimizar a generalização do modelo. Na avaliação final, no conjunto de testes, o modelo obteve uma acurácia de 99,52% e uma perda (loss) de 0,0488. Além disso, a inferência do modelo teve um tempo médio de execução de 63,2 us em CPU, evidenciando sua viabilidade para aplicações em tempo real. A matriz de confusão obtida é apresentada na Fig. 3. Nessa matriz, os valores 0, 1, 2 e 3 correspondem, respectivamente, as classes takeoff, land, stop e tracking, a relação é mostrada na Tabela I.

Para avaliar o desempenho, foram testados quatro gestos principais: palma aberta para ativar o tracking, mão fechada para stop, gesto de "V" para decolagem e gesto de "L" para pouso. Cada gesto foi avaliado em diversas condições e repetido múltiplas vezes por diferentes usuários, garantindo uma análise robusta da confiabilidade do reconhecimento. O

TABLE I
DESCRIÇÃO DAS CLASSES

Rótulo	Comando
0	Takeoff
1	Land
2	Stop
3	Tracking

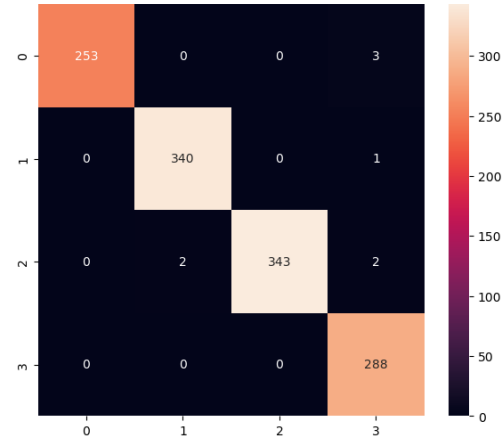


Fig. 3. Matriz de confusão.

resultado obtido da detecção desses gestos é apresentado na Fig. 4.

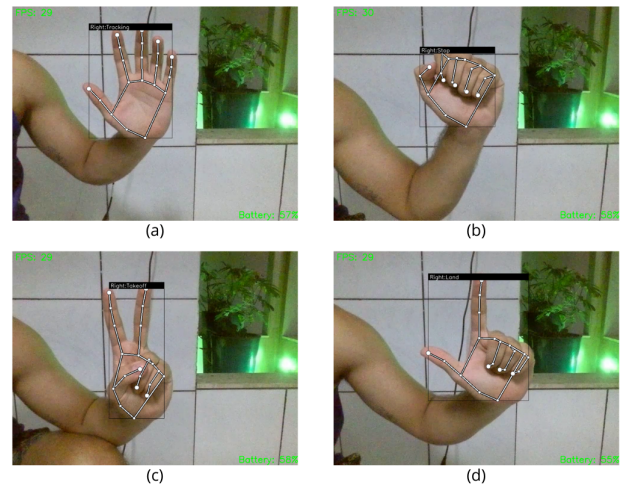


Fig. 4. Resultado da detecção.

Contudo, em ambientes com muitas pessoas, o método torna-se menos confiável, pois o MediaPipe pode se confundir na detecção de mãos, identificando, por vezes, uma mão mais distante da câmera do drone. Como o algoritmo não foi adaptado para lidar com essa situação, a detecção incorreta pode resultar no envio de comandos indesejados ao drone.

V. CONCLUSÃO

O sistema apresentou um desempenho satisfatório, no entanto, algumas limitações foram observadas. Condições de iluminação inadequadas afetaram a detecção da mão pelo MediaPipe, bem como o sistema de orientação do drone. Além disso, oclusões parciais da mão reduziram a confiabilidade do sistema em algumas situações, uma vez que é necessário que o usuário esteja sempre no campo de visão da câmera. Movimentos muito rápidos e abruptos também causaram erros na detecção, evidenciando a necessidade de um melhor tratamento dos gestos. Outro fator é a ausência de tratamento para exceções na classificação dos gestos. Isso significa que, se um gesto diferente dos treinados for apresentado ao modelo, ele será classificado como um dos gestos já aprendidos.

Em comparação com o controle por meio de dispositivos físicos, via aplicativo oficial do DJI Tello, o controle tradicional apresentou maior precisão para a execução de manobras mais complexas do que o controle por gestos, evidenciando que o método baseado em visão computacional é mais adequado para comandos básicos e interações simples e intuitivas.

A. Trabalhos Futuros

Alguns aspectos precisam de ajustes ou correções para viabilizar futuras melhorias e a implementação de novos recursos. Em especial, o controle de detecção de gestos deve ser aprimorado para aumentar a confiabilidade, reduzindo falhas na detecção, evitando confusões em caso de múltiplas mãos e garantindo um tratamento adequado para gestos não treinados.

Outro ponto de melhoria está relacionado ao drone utilizado. O DJI Tello foi escolhido por ser um drone comercial com programação aberta, o que elimina a necessidade de construir um drone do zero e facilita a implementação do programa. No entanto, ele apresenta limitações de hardware que afetam a precisão de seus movimentos. Para aplicações futuras que exijam maior precisão, será necessário optar por outro modelo de drone ou desenvolver um próprio.

Além disso, está em andamento uma investigação mais aprofundada para identificar outros cenários em que o método de controle apresenta falhas ou inconsistências.

Portanto, o projeto apresentado neste trabalho, aliado às melhorias propostas, pode contribuir significativamente para o avanço das pesquisas sobre o uso de inteligência artificial no controle de drones e veículos robóticos. Dessa forma, espera-se fomentar o uso inteligente de recursos e a integração de diversas tecnologias, impulsionando novas aplicações e inovações na área.

REFERÊNCIAS

- [1] T. W. Rauber, "Redes neurais artificiais," *Universidade Federal do Espírito Santo*, vol. 29, p. 39, 2005.
- [2] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.
- [3] K. Natarajan, T.-H. D. Nguyen, and M. Mete, "Hand gesture controlled drones: An open source library," in *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pp. 168–175, 2018.
- [4] D. S. CASANOVA, P. L. DE PAULA FILHO, K. SCHENATTO, and R. SOBJAK, "Reconhecimento de expressões faciais com mediapipe," in *Congresso Latino-Americano de Software Livre e Tecnologias Abertas (LATINOWARE)*, (Foz do Iguaçu/PR), pp. 257–263, Sociedade Brasileira de Computação, 2024.
- [5] P. M. Hell and P. J. Varga, "Drone systems for factory security and surveillance," *Interdisciplinary Description of Complex Systems: INDECS*, vol. 17, no. 3-A, pp. 458–467, 2019.
- [6] A. Sarkar, K. A. Patel, R. Ganesh Ram, and G. K. Kapoor, "Gesture control of drone using a motion controller," in *2016 International Conference on Industrial Informatics and Computer Systems (CIICS)*, pp. 1–5, 2016.
- [7] J.-W. Lee and K.-H. Yu, "Wearable drone controller: Machine learning-based hand gesture recognition and vibrotactile feedback," *Sensors*, vol. 23, no. 5, 2023.
- [8] S. Hadri, *HAND GESTURES FOR DRONE CONTROL USING DEEP LEARNING*. PhD thesis, 12 2018.
- [9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015.
- [10] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3d convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1–7, 2015.
- [11] K. Takahashi, "Hand gesture recognition using mediapipe." GitHub, 2021. [Online]. Disponível: <https://github.com/Kazuhito00/hand-gesture-recognition-using-mediapipe>. Acessado em: 10 mar. 2025.
- [12] R. V. F. d. e. a. SOUZA, "Drone-based object tracking with convolutional neural networks," in *Congresso Brasileiro Interdisciplinar em Ciência e Tecnologia*, (Diamantina(MG) Online), 2024. Acesso em: 10/03/2025.
- [13] Z. Drones, "Hand gesture control with tello drone," 2024. [Online]. Disponível em: <https://www.youtube.com/watch?v=d0Iza5IF7jE>. Acessado em: 12 mar. 2025.