Human Factors in Robots, Drones and Unmanned Systems, Vol. 218, 2026, 50–58
https://doi.org/10.54941/ahfe1007677

A Computer-Vision Approach to Accessible Robot Control: Hand Gesture Recognition for Users With Limited Mobility or Speech

Amin Majd, Mehdi Asadi, and Juha Kalliovaara

Turku University of Applied Sciences, Turku, 20520, Finland

ABSTRACT

Human–robot interaction increasingly demands intuitive, efficient, and accessible control mechanisms, particularly for users with physical or communication disabilities. Traditional interfaces—such as joysticks, keyboards, or voice commands—often impose significant cognitive or physical effort and may be unusable for individuals with impaired speech, hearing, or motor abilities. Recent advances in artificial intelligence and computer vision offer promising alternatives by enabling robots and autonomous systems to interpret human intentions directly from visual cues. This paper introduces a vision-based control framework that allows users to operate an autonomous drone through predefined hand gestures without any physical contact with a controller. The proposed system integrates real-time computer vision with control-system engineering to translate finger poses captured by a camera into actionable navigation commands. Our method employs PoseNet for robust hand-keypoint detection, combined with a custom gesture-classification module optimized for low-latency inference. The generated gesture classes are mapped to drone control instructions, enabling tasks such as takeoff, landing, directional movement, and hovering. The development process involved coordinated work across three subsystems: (1) Data Labeling, including dataset creation and annotation using CVAT and MATLAB; (2) Robot Interface and Connectivity, focusing on reliable communication between the vision module and the drone's flight controller; and (3) AI Model Development, comprising model selection, training, and optimization using Python, OpenCV, TensorFlow, and Google Colab. Although the project encountered initial technical and organizational challenges, the iterative development cycle ultimately led to a stable, functional prototype. Experimental results demonstrate that the system can accurately recognize gesture commands in real time and maintain responsive drone control under various lighting and background conditions. The achieved performance highlights the feasibility of replacing traditional physical controllers with AI-driven gesture interfaces, providing an accessible alternative for users who cannot operate conventional input devices. Overall, this work contributes a practical and innovative solution for enhancing human–robot interaction through contact-free control. The presented framework has potential applications not only in assistive technologies but also in fields such as rescue operations, manufacturing, and interactive robotics, where intuitive and hands-free control is advantageous. The project also offered valuable interdisciplinary experience in computer vision, robotics, and software engineering, demonstrating the effectiveness of merging AI-based perception with control-system design.

Keywords: Human factors in robots, Drones and unmanned systems

Received April 9, 2026; Revised May 9, 2026; Accepted May 29, 2026; Available online July 20, 2026

INTRODUCTION

Robots and autonomous agents have rapidly transitioned from confined industrial environments to becoming integral components of daily life, playing crucial roles in healthcare, logistics, and personal assistance (Goodrich and Schultz, 2007). By automating repetitive tasks, enhancing precision in complex operations, and providing reliable support in hazardous environments, these systems significantly improve both the quality of human life and overall operational efficiency (Murphy, 2000).

Despite the increasing autonomy of these systems, human-in-the-loop control remains essential to ensure safety, adaptability, and ethical oversight in dynamic, unpredictable scenarios (Endsley, 2017). As artificial intelligence continues to evolve, it empowers human operators with advanced support systems, scaling human capability by allowing a single operator to supervise and control entire swarms or coordinated groups of robots through unified, intuitive interfaces (Kolling et al., 2015).

Traditionally, human–robot interaction (HRI) has relied heavily on physical input devices, such as joysticks, keyboards, remote controllers, or, more recently, voice-command systems (Yanco and Drury, 2004). While effective for the general population, these conventional approaches often impose significant physical and cognitive demands. For individuals with motor impairments, limited mobility, or speech and hearing disabilities, manipulating a standard joystick with precision or issuing clear verbal commands can be physically exhausting or entirely impossible (Kiselev and Loutfi, 2012). Consequently, these traditional interfaces create a severe accessibility barrier. As industries increasingly integrate robotics into their daily workflows, this lack of accessible control mechanisms threatens to systematically exclude individuals with physical disabilities from a growing sector of the modern job market, depriving them of emerging employment opportunities and technological empowerment (Borg et al., 2011).

Addressing this critical gap requires a paradigm shift in how we design control interfaces, moving away from physically demanding hardware toward adaptive, software-driven solutions. Recent advances in artificial intelligence and computer vision present a promising avenue to bridge this accessibility divide. By leveraging AI to interpret human intentions directly from natural, non-verbal visual cues, we can develop intuitive systems that bypass the physical limitations of the user, transforming inherent human movements into precise machine commands (Rautaray and Agrawal, 2015).

Building upon this premise, this paper introduces a novel, vision-based control framework that enables users to operate an autonomous drone entirely hands-free and without any physical contact with a controller. Our proposed system utilizes real-time computer vision, specifically employing PoseNet for robust hand-keypoint detection, paired with a custom low-latency gesture-classification module. By translating predefined finger poses into actionable navigation commands—such as takeoff, landing, hovering, and directional movement—the system establishes a seamless communication channel between the user and the drone's flight controller. This framework was developed through a rigorous process involving custom data labeling

with CVAT, AI model optimization using TensorFlow, and the establishment of reliable interface connectivity. Experimental results validate that this AI-driven approach provides high accuracy and responsiveness under varied lighting conditions, offering a highly practical, accessible alternative to conventional physical controllers.

The remainder of this paper is organized as follows: Section II reviews the background and related work in accessible HRI and vision-based control. Section III details the proposed approach, including system architecture, data processing, and AI model development. Section IV presents the experimental results and performance evaluation. Finally, Section V concludes the paper and discusses future research directions and applications.

BACKGROUND AND REKATED WORKS

The development of accessible human–robot interaction (HRI) systems exists at the intersection of control engineering, assistive technology, and computer vision. This section reviews the evolution of robot control interfaces, the limitations they impose on users with disabilities, and the emergence of AI-driven vision models as a viable solution.

Conventional Control Interfaces and Accessibility Barriers

The dominant paradigm for controlling unmanned aerial vehicles (UAVs) and ground robots heavily relies on handheld Radio Frequency (RF) controllers, joysticks, or keyboard-and-mouse setups (Valavanis and Vachtsevanos, 2015). These traditional physical interfaces are engineered for users with typical physical capabilities, requiring high degrees of manual dexterity, fine motor control, bilateral hand coordination, and sustained grip strength. More recently, voice-user interfaces (VUIs) have been integrated into autonomous systems to provide a "hands-free" alternative (Marge et al., 2022).

However, these conventional modalities present profound accessibility barriers. For individuals with upper-limb motor impairments—resulting from conditions such as spinal cord injuries, cerebral palsy, muscular dystrophy, or amputation—manipulating physical joysticks is either physically exhausting or entirely unfeasible (Wobbrock et al., 2011). Furthermore, individuals experiencing speech impairments (e.g., dysarthria, aphasia) or those operating in noisy industrial environments cannot reliably use voice-command systems, as these systems often fail to accurately parse non-standard speech patterns (Hawley et al., 2007). Consequently, reliance on these standard interfaces inherently excludes a significant demographic from interacting with modern robotic systems, necessitating alternative control methodologies.

Assistive Approaches in Human-Robot Interaction

To address these limitations, researchers have explored various assistive technologies for HRI. Existing solutions generally fall into two categories: specialized mechanical switches and biosignal-based interfaces. Mechanical assistive devices, such as sip-and-puff systems or head-array switches, allow

users to issue commands using minimal movement or breath (Simpson, 2005). While reliable, these systems suffer from extremely low bandwidth; complex tasks like drone navigation require multiple sequential switch activations, resulting in high cognitive load and slow response times.

Alternatively, biosignal interfaces such as Brain-Computer Interfaces (BCIs) (Millan et al., 2004) and electromyography (EMG) armbands (Wolf et al., 2013)attempt to decode user intent directly from neurological or muscular activity. Although BCIs offer a hands-free approach, they typically require users to wear intrusive sensor caps, necessitate tedious calibration, and are susceptible to signal noise. Similarly, gaze-tracking interfaces (Jacob, 1990) have been proposed for drone control, but they often suffer from the "Midas touch" problem, where the system struggles to differentiate between casual observation and intentional command issuing. Therefore, a critical need remains for a high-bandwidth, non-intrusive interface that does not require the user to wear specialized hardware.

Computer Vision and Human Pose Estimation

The rapid advancement of deep learning and computer vision has introduced highly promising, non-contact solutions for HRI. By utilizing standard RGB cameras, modern AI frameworks can accurately interpret human spatial features and postures, translating them into control signals without requiring physical sensors on the user's body (Shotton et al., 2011).

Early vision-based gesture recognition systems relied heavily on depth-sensing cameras (e.g., Microsoft Kinect) or infrared sensors to isolate the user from the background (Ren et al., 2011). However, the advent of Convolutional Neural Networks (CNNs) has enabled robust Human Pose Estimation (HPE) directly from standard 2D video feeds. State-of-the-art frameworks like OpenPose (Cao et al., 2019) and MediaPipe (Lugaresi et al., 2019) have demonstrated exceptional accuracy in mapping the human skeleton and hand joints in real time.

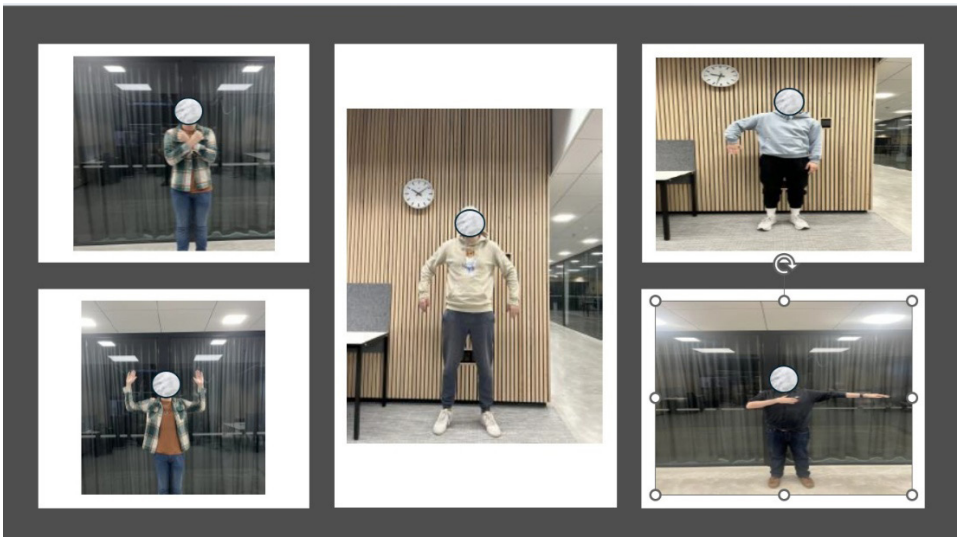

Figure 1: Samples of different poses that the AI model has learned.

Among these architectures, PoseNet has emerged as a highly efficient model for real-time posture translation. PoseNet utilizes a lightweight CNN backbone (often MobileNet or ResNet) to predict the spatial coordinates of key anatomical landmarks (keypoints) on the human body, such as wrists, elbows, and finger joints (Papandreou et al., 2017). Unlike object-detection models (such as YOLO (Redmon et al., 2016)) that merely draw bounding boxes around a hand, keypoint-based models like PoseNet extract the precise geometric configuration of the hand.

This level of granularity is crucial for accessible control. By capturing the exact positional data of the fingers and palm, PoseNet allows for the algorithmic classification of subtle, predefined hand gestures (Zhang et al., 2020) Because it is optimized for low-latency inference, PoseNet can continuously process video streams on edge devices without relying on massive cloud-computing resources. Our proposed framework builds upon these computer vision principles, utilizing real-time keypoint extraction to bypass physical limitations and map intuitive hand postures directly to drone flight kinematics.

PROPOSED WORK

The primary objective of this research is to establish a robust, contact-free control framework that translates human spatial gestures into real-time unmanned aerial vehicle (UAV) kinematics. To achieve this, we developed a closed-loop pipeline encompassing computer vision, machine learning, and control-system engineering. The framework is deployed on a DJI Robomaster TT drone, utilizing its onboard camera for visual perception, its flight controller for navigation, and an expansion kit featuring a programmable LED dot-matrix display for visual feedback.

Our proposed architecture is modularly divided into three sequential subsystems: (1) Data Collection and Spatial Normalization, (2) AI Model Development, and (3) Real-Time Inference and Kinematic Mapping.

Pose Extraction and Spatial Normalization

The foundation of our gesture-recognition system relies on the accurate extraction of anatomical landmarks from 2D RGB image data. Rather than relying on computationally heavy pixel-based image classification, our approach extracts a skeletal wireframe using Google's MediaPipe framework, which excels in low-latency Human Pose Estimation (HPE).

To construct the training dataset, varied images of the operator executing seven predefined navigational gestures (Up, Down, Left, Right, Forward, Backward, and Stop) were captured. The HPE algorithm isolates critical anatomical keypoints in 3D space (x, y, z), specifically tracking the nose, shoulders, elbows, wrists, and hips.

A critical challenge in vision-based control is ensuring scale and translation invariance. The system must recognize a gesture regardless of the operator's distance from the camera or their specific body proportions. To solve this, we implemented a robust mathematical normalization algorithm during the

data preprocessing stage. For every frame, the algorithm calculates a dynamic reference anchor centered on the operator's hips. The spatial coordinates of all upper-body landmarks are then recalculated relative to this central anchor point. Furthermore, the coordinate values are divided by the calculated torso length (the vertical distance between the hip center and the nose).

This torso-length normalization ensures that a gesture performed by a tall operator close to the camera outputs the exact same numerical data array as the same gesture performed by a shorter operator standing further away. Frames lacking sufficient torso visibility are algorithmically discarded to prevent training data corruption.

Figure 2: Testing the approach in a real enviroments.

Data Augmentation and Model Training

To maximize the generalization capabilities of the neural network and prevent overfitting, artificial data augmentation techniques were programmatically applied to the normalized dataset. Using OpenCV's affine transformation functions, the base dataset was multiplied by introducing controlled randomized variations. Each base image generated multiple augmented copies featuring random rotational shifts between –15 and +15 degrees, as well as synthetic scaling adjustments ranging from 90% to 110% of the original size.

The augmented, normalized coordinate arrays were labeled according to their corresponding navigational commands and exported to a structured dataset. This dataset was systematically partitioned into a 70% training subset and a 30% testing/validation subset. A lightweight, multi-class neural network classifier was constructed using TensorFlow and Scikit-learn. By training exclusively on the normalized spatial coordinates rather than raw pixel arrays, the model was heavily optimized for rapid inference, successfully learning the non-linear boundaries between the seven distinct gesture classes with high precision.

Real-Time Inference and Confidence Filtering

During live operation, the system establishes a continuous video feed via the drone's onboard camera or an auxiliary webcam. Frame by frame, the video stream undergoes the identical extraction and normalization pipeline established during the training phase. The resulting live coordinate arrays are fed into the optimized machine learning model, which outputs a predicted gesture class alongside a probabilistic confidence score.

To ensure operational safety and prevent erratic drone behavior caused by transitional human movements (e.g., the operator naturally lowering their arms), we implemented a dynamic confidence-filtering mechanism. Gestures are only translated into execution commands if the model's classification confidence exceeds predefined thresholds. While standard directional movements (such as Up, Down, Left, Right) require a baseline confidence of >75% to trigger, high-risk maneuvers—such as the "Forward" command, which reduces the distance between the UAV and the user—are restricted by a stringent 99% confidence threshold. This dual-tier filtering drastically reduces the false-positive rate.

Kinematic Mapping and Multi-Modal Control

Once a gesture passes the confidence filter, the classification is passed to a unified Python control script that translates the gesture into localized drone kinematics. The system acts as a direct interface to the drone's flight controller. For example, a recognized "Stop" pose zeroes out all velocity vectors, forcing the drone into a stable hover. Directional gestures trigger parameterized spatial displacement (e.g., executing a 20 cm translation in the specified vector).

Furthermore, the framework expands beyond basic gesture navigation by introducing multi-modal AI tracking. We engineered a specific "Switch Pose" tied to distinct hip-movement tracking. Upon recognizing this specific toggle sequence, the core script suspends the MediaPipe Pose logic and seamlessly transitions the drone into an active Face Recognition and Tracking mode utilizing OpenCV Haar cascades or deep-learning-based face trackers. In this mode, the drone autonomously rotates on its yaw axis to continuously keep the operator's face centered in the frame. The operator can seamlessly toggle back to Pose Control via the designated switch gesture.

Finally, state-feedback is communicated to the user via the drone's LED dot-matrix module. The expansion module emits specific color codes associated with the current active mode (Pose Control vs. Face Tracking), providing crucial visual confirmation to the operator and closing the human-robot interaction loop entirely without physical hardware.

CONCLUSION AND FUTURE WORKS

As robotic and autonomous systems become increasingly integrated into everyday life and industrial operations, the reliance on physically demanding control interfaces remains a critical barrier for users with limited mobility or speech impairments. To address this challenge, this paper presented

a novel, contact-free control framework that translates human spatial gestures into real-time drone kinematics using advanced computer vision. By leveraging keypoint extraction, dynamic spatial normalization, and a lightweight machine learning classifier, we successfully replaced traditional physical hardware with an intuitive, AI-driven interface. Our experimental implementation demonstrated robust real-time responsiveness, accurately executing navigational commands and multi-modal tracking without requiring the user to wear specialized sensors.

Ultimately, this research provides a practical proof of concept for bridging the accessibility gap in human–robot interaction. While the current prototype successfully manages fundamental flight mechanics, it represents a crucial starting point for broader assistive applications. Future extensions of this work aim to integrate finer hand-articulation tracking, adaptive learning algorithms personalized to an individual's specific mobility constraints, and autonomous obstacle avoidance to further reduce cognitive load. By continuing to refine these vision-based frameworks, the robotics community can ensure that emerging technologies empower, rather than exclude, individuals with physical disabilities, fostering a more inclusive technological landscape and opening new avenues for independence, accessibility, and employment.

REFERENCES

Goodrich, M. A., & Schultz, A. C. (2007). Human-robot interaction: A survey. Foundations and Trends in Human-Computer Interaction, 1(3), 203-275. Hanover, MA: Now Publishers.

Murphy, R. R. (2000). Introduction to AI robotics. Intelligent Robotics and Autonomous Agents Series. Cambridge, MA: MIT Press.

Endsley, M. R. (2017). Toward a theory of situation awareness in dynamic systems. Situation Awareness Analysis and Measurement. Boca Raton, FL: CRC Press.

Kolling, A., Walker, P., Chakraborty, N., Sycara, K., & Lewis, M. (2015). Human interaction with robot swarms: Survey, taxonomy, and future directions. IEEE Transactions on Human-Machine Systems, 46(1), 9-26. Piscataway, NJ: IEEE.

Yanco, H. A., & Drury, J. (2004). Classifying human-robot interaction: An updated taxonomy. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. Piscataway, NJ: IEEE.

Kiselev, A., & Loutfi, A. (2012). Using a gesture interface for teleoperation of a mobile robot. Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction. New York, NY: ACM.

Borg, J., Larsson, S., & Östergren, P. O. (2011). The right to assistive technology: For whom, for what, and by what means?. Disability & Society, 26(2), 151-167. Abingdon, UK: Routledge.

Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: A survey. Artificial Intelligence Review, 43(1), 1-54. Dordrecht, Netherlands: Springer.

Valavanis, K. P., & Vachtsevanos, G. J. (2015). Handbook of unmanned aerial vehicles. Dordrecht, Netherlands: Springer.

Marge, M., Bonial, C., Byrne, B., Cassidy, T., Hudson, A., Hayes, C. J., & Traum, D. (2022). Spoken language interaction with robots: Research issues and directions. Computer Speech & Language, 71, 101232. Amsterdam, Netherlands: Elsevier.

Wobbrock, J. O., Kane, S. K., Gajos, K. Z., Harada, S., & Froehlich, J. (2011). Ability-based design: Concept, principles and examples. ACM Transactions on Accessible Computing, 3(3), 1–27. New York, NY: ACM.

Hawley, M. S., Enderby, P., Green, P., Cunningham, S., Brownsell, S., Carmichael, J., & O'Neill, P. (2007). A speech-controlled environmental control system for people with severe dysarthria. Medical Engineering & Physics, 29(5), 586–593. Amsterdam, Netherlands: Elsevier.

Simpson, R. C. (2005). Smart wheelchairs: A literature review. Journal of Rehabilitation Research & Development, 42(4), 423–436. Washington, DC: Department of Veterans Affairs.

Millan, J. D. R., Renkens, F., Mouriño, J., & Gerstner, W. (2004). Noninvasive brain-actuated control of a mobile robot by human EEG. IEEE Transactions on Biomedical Engineering, 51(6), 1026–1033. Piscataway, NJ: IEEE.

Wolf, M. T., Assad, C., Vernacchia, M. T., Fromm, J., & Jethani, H. L. (2013). Gesture-based robot control with an electromyography armband: A feasibility study. Proceedings of the IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE.

Jacob, R. J. (1990). What you look at is what you get: Eye movement-based interaction techniques. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY: ACM.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., & Blake, A. (2011). Real-time human pose recognition in parts from single depth images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE.

Ren, Z., Yuan, J., Meng, J., & Zhang, Z. (2011). Robust part-based hand gesture recognition using kinect sensor. IEEE Transactions on Multimedia, 15(5), 1110-1120. Piscataway, NJ: IEEE.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1), 172–186. Piscataway, NJ: IEEE.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., & Grundmann, M. (2019). MediaPipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172. Ithaca, NY: Cornell University.

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., & Murphy, K. (2017). Towards accurate multi-person pose estimation in the wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE.

Zhang, X., Yang, Z., Feng, T., Zheng, Y., & Hou, X. (2020). Hand gesture recognition for drone control using OpenPose. Proceedings of the IEEE 3rd International Conference on Artificial Intelligence and Big Data. Piscataway, NJ: IEEE.