Original Article
ISSN (Online): 2454-1907

International Journal of Engineering Technologies and Management Research
May 2025 12(5), 1–10

TWO-HANDED DYNAMIC GESTURE RECOGNITION USING RGB-D SENSORS

Yu-Chi Pu [1], Wei-Chang Du [2] ✉, Kai-Wei Shih [2]

[1] Department of Maritime Information and Technology, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan, Republic of China
[2] Department of Information Engineering, I-Shou University, Kaohsiung, Taiwan, Republic of China

Received 05 March 2025
Accepted 02 April 2025
Published 07 May 2025

CorrespondingAuthor

Wei-Chang Du, wcdu@isu.edu.tw

DOI
10.29121/ijetmr.v12.i5.2025.1568

Funding: This research is partly supported by the Ministry of Science and Technology of Taiwan under Grant MOST110-2221-E-214-020.

ABSTRACT

Sign language is a form of visual-gestural communication that conveys semantic content through hand movements and postures. It comprises a structured set of gestures, each associated with specific meanings. For individuals with hearing impairments, sign language serves as a primary medium for expression, perception, and interaction with the external world. However, due to the general lack of sign language proficiency among the broader population, effective communication remains a significant challenge. This paper uses an RGB-D imaging device to capture both color and depth information of dynamic hand gestures, enabling more robust and discriminative feature extraction than traditional RGB-based approaches. The proposed system focuses on recognizing two-handed dynamic gestures by analyzing spatial configurations and temporal motion patterns. A gesture symbolization mechanism facilitates the recognition process, wherein complex gesture sequences are encoded into a set of primitive symbols representing key postures and transitions. These symbolic representations are then compared using a fuzzy matching algorithm, which accounts for variations in gesture execution and temporal alignment, thereby enhancing the system's tolerance. This methodology aims to provide a reliable and flexible framework for real-time gesture recognition in natural interaction scenarios.

Keywords: Human-Computer Interaction, RGB-D Sensors, Gesture Recognition

1. INTRODUCTION

In recent years, affordable depth-sensing technologies have spurred a new wave of development in Natural User Interfaces (NUIs). Depth sensors such as Intel RealSense and Microsoft Kinect generate motion capture data that provide spatial and temporal descriptions of human skeletal structures, typically represented as hierarchical models composed of key joints. Concurrently, the applications of RGB-D imaging have expanded significantly, supporting a wide range of domains including robot vision, 3D reconstruction, video surveillance, and motion tracking systems Shaikh and Chai (2021). By combining real-world depth information with

How to cite this article (APA): Pu, Y. C., Du, W. C., and Shih, K. W. (2025). Two-Handed Dynamic Gesture Recognition Using RGB-D Sensors. *International Journal of Engineering Technologies and Management Research, 12*(5), 1–10. doi: 10.29121/ijetmr.v12.i5.2025.1568

RGB data, RGB-D imaging enables richer scene understanding and more immersive interactive experiences.

One of the most impactful applications of RGB-D data lies in its ability to enable intuitive, device-free human-computer interaction (HCI). By estimating skeletal postures in real-time, user movements can be interpreted without the need for handheld controllers or wearable sensors. This capability has accelerated the development of interaction systems in fields such as sign language translation, virtual reality, and assistive technologies. However, due to inherent limitations such as occlusions, limited spatial resolution, and background interference, consumer-grade RGB-D sensors often produce noisy or incomplete data, particularly when capturing fine-grained dynamic hand gestures. These challenges highlight the need for more robust and adaptive gesture recognition techniques.

1.1. DEPTH SENSING TECHNOLOGY

Early research on posture and motion capture originated in the field of biomechanics and has, over the past several decades, been increasingly applied to domains such as human-computer interaction, computer animation, and virtual reality. These applications primarily involve recording and analyzing human movement in digital form. Traditional motion capture systems often required physical sensors to be attached to key joints of the body, transmitting data to computers for real-time processing. To address the needs of interactive systems in areas such as entertainment, healthcare, training, and smart home environments, various motion capture technologies have been developed, including mechanical, electromagnetic, inertial, and optical systems. Optical motion capture can be further categorized into marker-based and markerless approaches. Marker-based systems typically offer higher accuracy but are expensive and require users to wear specialized equipment. In contrast, markerless systems offer a more cost-effective and user-friendly alternative. Although markerless systems may yield slightly lower precision, their convenience and affordability make them well-suited for widespread adoption in real-world applications.

In this paper, the Intel RealSense SR300 depth-sensing camera Keselman et al. (2017) is employed as the primary sensing device. It utilizes a structured light technique, in which near-infrared (IR) laser patterns are projected onto the scene. The reflected IR light is captured by a dedicated lens and decoded by an onboard processing unit to compute per-pixel depth information. Given the wide variability in possible hand gestures, the extraction of skeletal data from the depth image serves as a critical foundation for gesture recognition. The RealSense is a short-range, high-resolution depth camera that is particularly well-suited for capturing fine hand movements. In this paper, the hand tracking module provided by the RealSense SDK is used to capture and recognize two-handed dynamic gestures. Specifically, the SDK enables the detection of 22 hand landmarks per hand, including four joints for each finger and additional key points on the palm and wrist. Based on the trajectories and configurations of these landmarks, this study constructs a framework for two-handed dynamic gesture recognition.

1.2. MOTIVATION AND PURPOSE

In recent years, RGB-D motion sensing devices have gained significant attention due to their markerless optical sensing capabilities. As a cost-effective and non-intrusive solution, RGB-D technology has become a mainstream approach in motion-sensing applications. Unlike conventional RGB cameras, RGB-D devices provide both color and depth information, thereby enhancing visual perception and spatial understanding in human-computer interaction systems.

The affordability and ease of use of depth sensing technologies have accelerated their adoption in consumer and industrial applications, facilitating the development of intuitive and accessible HCI solutions. Despite these advantages, RGB-D sensors continue to face several technical limitations that hinder their effectiveness in capturing reliable motion data. Some of the major challenges include:

- **Self-occlusion:** During gesture execution, parts of the body may occlude one another from the sensor's viewpoint, resulting in incomplete or inaccurate skeletal reconstruction.
- **Depth discontinuities and image holes:** Highly reflective or transparent surfaces (e.g., glass or mirrors) can lead to missing depth information. In addition, due to the spatial offset between the infrared projector and receiver, occlusion artifacts and boundary holes often occur, especially around foreground-background transitions.
- **Depth noise:** Environmental infrared interference and sensor limitations may introduce random noise patterns into the depth image, adversely affecting gesture segmentation and feature extraction.
- **Rapid motion artifacts:** Sudden changes in velocity or direction during gesture execution may exceed the sensor’s temporal resolution, leading to erratic or discontinuous skeletal tracking.

This paper addresses these challenges by focusing on the real-time recognition of two-handed dynamic gestures using RGB-D input under high-noise conditions. Dynamic gestures are composed of a sequence of static poses, each of which can be represented by the spatial configuration of multiple skeletal joints. As such, dynamic gestures inherently encode both spatial and temporal information. Spatially, joint positions exhibit a hierarchical structure with varying degrees of freedom and rotation constraints. Temporally, meaningful motion patterns emerge from the sequential dependency between consecutive gestures Osman et al. (2024). In contrast to conventional static gesture recognition or single-hand dynamic gesture analysis, this paper emphasizes the recognition of complex two-handed dynamic gestures, even under conditions with noisy or incomplete depth data. The proposed approach aims to provide a robust and practical solution for gesture recognition in real-world environments. The system captures spatial-temporal features from hand motion sequences and applies a gesture symbolization process that encodes key postures into string representations. To improve tolerance to individual variation and temporal inconsistency, some fuzzy matching algorithms are employed to compare gesture patterns and enhance recognition accuracy. The proposed framework aims to advance the development of real-time, accurate, and user-friendly gesture-based interaction systems, particularly for applications in sign language communication and naturalistic human-computer interaction.

2. RELATED WORK

With the growing prevalence of RGB-D sensing technologies, dynamic gesture recognition has attracted increasing attention in recent years. The integration of depth sensing effectively addresses the long-standing challenge of background interference, thereby enabling natural human gestures to emerge as a promising modality for human-computer interaction Yasen and Jusoh (2019). Traditionally, HCI has relied on physical input devices such as the mouse, keyboard, and touch panel. To achieve more intuitive interactions, removing these physical constraints has become a key direction of development. Among various alternatives, gesture recognition remains a highly promising area of research. In early approaches, gesture information was often acquired through data gloves, which transmitted motion data to computers for the recognition of static gestures Lei et al. (2015). Although data gloves provide high precision and are less affected by lighting conditions compared to conventional cameras, they are expensive and cumbersome to use, limiting their scalability and practicality in real-world applications.

RGB images, on the other hand, can be acquired through widely available devices such as webcams, smartphones, laptops, and tablets. Since these methods do not require users to wear additional sensors, gesture recognition via image analysis is anticipated to become a mainstream approach in future HCI systems Parveen et al. (2020). One commonly adopted technique involves detecting hand regions based on skin color segmentation. This approach typically employs decision-tree algorithms to identify the skin color range and then locates corresponding regions in the image. However, color-based methods are highly susceptible to variations in ambient lighting, background complexity, and shadows, often resulting in reduced tracking accuracy.

In contrast, depth images encode distance information within a scene, enabling the extraction of relative spatial relationships between pixels. This facilitates efficient foreground-background segmentation and typically requires less computational overhead compared to color-based methods, making depth imaging particularly well-suited for gesture recognition tasks Sun et al. (2023). The use of depth cameras provides improved speed and accuracy in gesture tracking, aligning well with the demands of modern HCI applications. Recent studies have explored the application of machine learning techniques to enhance gesture recognition performance. For example, some studies have proposed classification models utilizing both RGB and depth modalities, employing architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with long short-term memory (LSTM) units to capture spatial and temporal features in dynamic gestures Obaid et al. (2020), Rahman et al. (2025).

3. GESTURE ANALYSIS MODEL

This paper uses RGB-D images to analyze and recognize two-handed dynamic gestures. Initially, hand joint data from both hands are extracted to facilitate static gesture recognition based on distinctive gesture features. To address the substantial noise inherent in sensing devices, some fuzzy string-matching methods are employed to enhance the robustness of dynamic gesture recognition.

3.1. STATIC GESTURE FEARURE EXTRACTION AND RECOGNITION

Due to the inherent instability and limited precision of low-cost depth cameras, the accuracy of skeletal joint estimation is often compromised. Although various noise reduction techniques can alleviate this issue, minor oscillations may still persist in the resulting posture sequences. Nevertheless, the depth video data ensures that a small degree of noise does not significantly affect the analysis of overall gesture patterns.

In this paper, depth images are acquired using an Intel RealSense sensor. A trained classifier is employed to automatically segment hand regions at the pixel level, after which the spatial positions of hand joints are estimated using the mean shift algorithm. The skeletal model includes 22 key joint points: four nodes for each finger (thumb, index, middle, ring, and little finger), and one node each for the wrist and palm. These joints are labeled sequentially as J_0 through J_{21}, as illustrated in Figure 1.

Figure 1

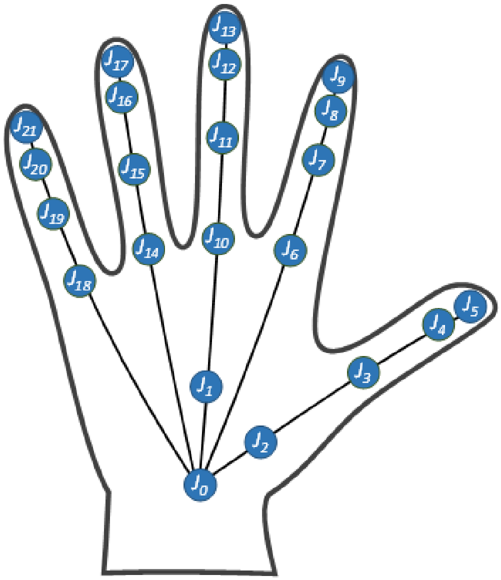

Figure 1 Schematic Diagram of Hand Joints

Effective description and extraction of gesture features are critical steps in tasks such as action recognition, data retrieval, and semantic interpretation. However, due to the subjective nature of human perception, designing features that align with human intuition remains challenging. Despite this, it is essential for gesture features to maintain invariance under spatial and temporal transformations to ensure robustness and generalizability. To accommodate diverse hand configurations, this study quantifies the degree of finger flexion by measuring the bending angles of finger joints (i.e., knuckle angles). These angles serve as key features for distinguishing between various open and closed hand poses. As illustrated in Figure 1, a five-dimensional feature vector is constructed using the following joint angles:

- Thumb: $\angle J_0J_2J_5$
- Index finger: $\angle J_6J_7J_9$
- Middle finger: $\angle J_{10}J_{11}J_{13}$
- Ring finger: $\angle J_{14}J_{15}J_{17}$
- Little finger: $\angle J_{18}J_{19}J_{21}$

A set of basic static gestures are predefined based on common gestures observed in daily life. Under the supervised learning framework, the K-nearest neighbor (KNN) algorithm is used to classify the static gestures of the left and right hands. Each recognized static gesture is then encoded using symbolic character labels for subsequent gesture sequence analysis.

3.2. DYNAMIC GESTURE FEATURE RECOGNITION

Currently, most common gesture recognition methods use single-hand gestures, and there are relatively few gestures that can be combined. This paper further considers using continuous two-handed gestures for related recognition. Dynamic gestures are composed of multiple static gestures. With the use of both hands, a variety of combinations can be created to achieve the diversity required for human-computer interaction. After obtaining continuous static gestures, the gestures are converted into corresponding strings, and then the string similarity is compared to identify the gesture action. This study uses fuzzy string-matching to calculate string similarity Rudwan and Fonou-Dombeu (2023) in order to make correct judgments under tolerable noise interference.

- **Metric-LCS:** The longest common subsequence metric (Metric-LCS for short) is an extension of the longest common subsequence (LCS). LCS is mainly a way to find the longest common subsequence between two strings. This sequence does not need to be a continuous position in the original sequence. Metric-LCS converts the LCS length into the similarity between two strings S_1 and S_2:

$$MLCS(S_1, S_2) = \frac{LCS(S_1,S_2)}{max(|S_1|,|S_2|)} \tag{1}$$

The obtained value range is [0,1], and the larger the value, the more similar the two strings are.

- **Jaccard Index:** Jaccard Paul proposed the Jaccard index in 1901. It mainly converts the strings S_1 and S_2 into sets X_1 and X_2 represented by single characters, and then calculates the similarity of the two sets:

$$Jaccard(X_1, X_2) = \frac{|X_1 \cap X_2|}{|X_1 \cup X_2|} \tag{2}$$

The Jaccard index calculates the ratio of the intersection of sets to the union of sets, and then estimates the similarity of sets.

- **Jaro-Winkler Similarity:** Jaro-Winkler Similarity was proposed by Winkler in 1990. It is an extension of Jaro Similarity. It can take into account the transposition of adjacent characters in the comparison process and is suitable for situations where the string length is short. During the matching process, the size of the matching window must be met:

$$MW = \frac{\max\,(|S_1|,|S_2|)}{2} - 1 \tag{3}$$

Jaro similarity is defined as follows:

$$Sim_j(S_1, S_2) = \frac{1}{3}\left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m-t}{m}\right) \tag{4}$$

where $Sim_j(S_1,S_2)$ is the similarity between strings S_1 and S_2, and the higher the value, the more similar they are. The symbol m represents the number of characters that are identical between the two strings and meet the MW distance, and t is half of the number of characters that need to be transposed when the characters obtained by m are compared with the target string. Jaro-Winkler similarity is an extended calculation based on Jaro similarity and is defined as follows:

$$Sim_{jw}(S_1, S_2) = Sim_j(S_1, S_2) + \ell \cdot p \cdot (1 - Sim_j(S_1, S_2)) \tag{5}$$

where l is the number of identical prefixes in the two strings, and p is the weight value for adjusting the common prefix character. The higher the calculated value, the more similar it is. When two strings have the same character at the beginning, they have a higher weight value.

By employing these similarity measures, the proposed system enhances the reliability of dynamic gesture recognition, even under imperfect sensor conditions.

4. EXPERIMENTAL RESULTS

In view of the increasing importance of cost-effective depth sensing technology in its potential for widespread applications, a series of experiments were conducted in this paper to evaluate the feasibility and effectiveness of the proposed method.

4.1. STATIC GESTURE RECOGNITION

To evaluate static gesture recognition performance, fifteen fundamental gestures were defined based on commonly used hand shapes in American Sign Language (ASL), as illustrated in Figure 2. For each gesture, 20 samples were collected, resulting in a total of 300 test samples. The K-nearest neighbors (KNN) algorithm was employed for classification, with the dataset evenly divided into training and testing sets. To determine the optimal K value, classification accuracy was evaluated across multiple values of K. As shown in Figure 3, the recognition accuracy reached its peak at K=3, achieving an overall accuracy of 84%. Notably, gesture P9, which corresponds to a fully open palm, achieved 100% accuracy under all tested conditions. In contrast, gestures P11 to P14, characterized by partially bent fingers and joint occlusion, exhibited lower recognition rates due to their less distinguishable joint configurations.

Figure 2

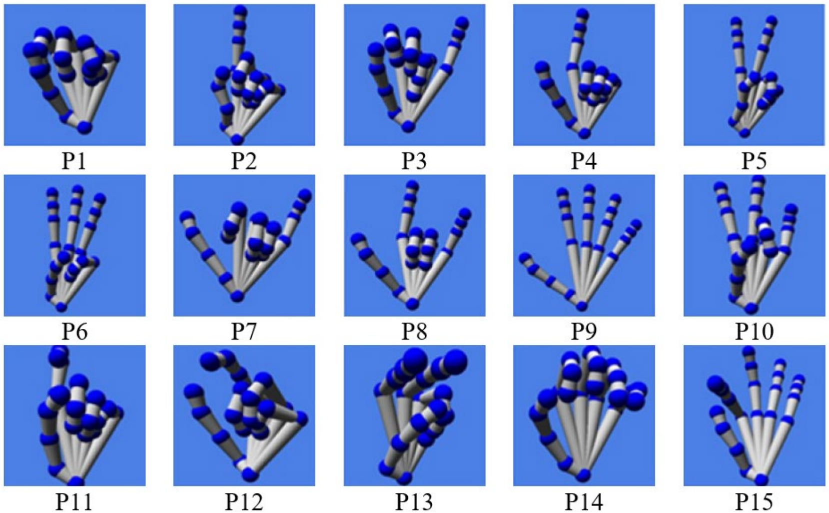

Figure 2 Static Gesture Diagram, Wu et al. (2020)

Figure 3

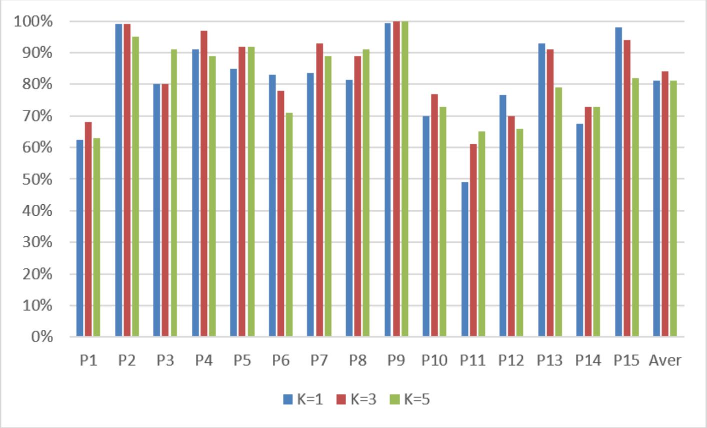

Figure 3 Accuracy of Static Gesture Recognition

4.2. DYNAMIC GESTURE RECOGNITION

To evaluate the performance of two-handed dynamic gesture recognition, this paper defines 10 distinct categories of dynamic gestures. Each gesture category comprises between 151 and 387 individual static gesture frames. For each category, ten gesture sequences were recorded, yielding a total of one hundred gesture sequence datasets. An overview of the defined gesture categories is provided in Figure 4. In the recognition pipeline, each static frame within a dynamic gesture sequence is first classified using the previously trained K-nearest neighbors static gesture recognizer. This yields a pair of class labels per frame. These labels are then encoded as single-character representations to form a symbolic string corresponding to the gesture sequence

The dynamic gesture recognition is subsequently performed using string similarity comparison. For each gesture category, one gesture sequence is randomly selected as the reference template. The remaining sequences are compared against these reference templates using string similarity metrics. The category associated with the highest similarity score is assigned as the predicted label for the tested sequence. A total of 100 recognition trials were conducted using this method. The experimental results, summarized in Figure 5, indicate that most similarity metrics achieve recognition accuracies exceeding 70%. Among them, the Metric-LCS method demonstrates the best overall performance, with an average accuracy of 79% across all gesture categories.

Figure 4

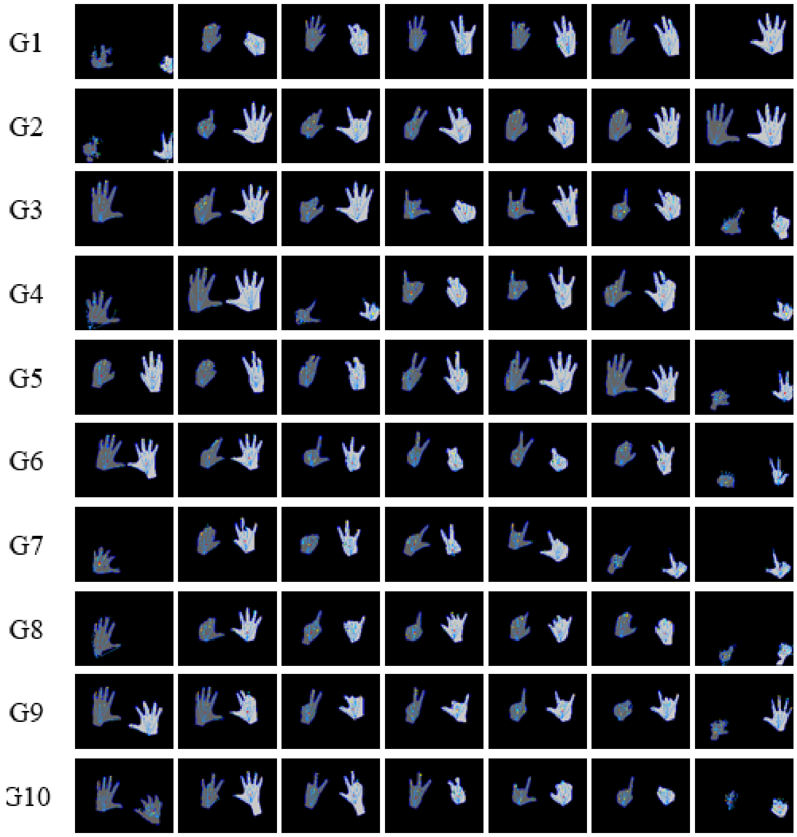

Figure 4 Dynamic Gesture Examples

Figure 5

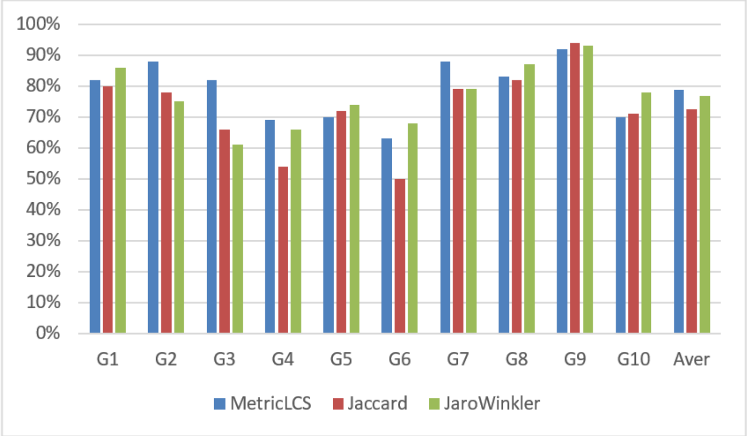

Figure 5 Accuracy of Dynamic Gesture Recognition

5. CONCLUSION

This paper presents a novel approach for recognizing both static and two-handed dynamic gestures using RGB-D images. Despite the inherent challenges of self-occlusion in hand postures, the proposed method achieves over 80% accuracy for static gestures and nearly 80% for dynamic gesture sequences. The experimental results demonstrate the feasibility and effectiveness of integrating skeletal joint estimation with fuzzy string-matching techniques for robust gesture analysis.

The framework developed in this work is designed for general-purpose gesture recognition; however, its performance can be further improved by incorporating task-specific gesture features tailored to specific application domains. Future research could focus on optimizing feature representations using contextual or semantic information to improve recognition accuracy and broaden applicability in human-computer interaction scenarios.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Keselman, L., Woodfill, J. I., Grunnet-Jepsen, A., & Bhowmik, A. (2017). Intel Realsense Stereoscopic Depth Cameras. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1267-1276. https://doi.org/10.1109/CVPRW.2017.167

Lei, L., Jinling, Z., Yingjie, Z., & Hong, L. (2015). A Static Gesture Recognition Method Based on Data Glove. Journal of Computer-Aided Design & Computer Graphics, 27(12), 2410-2418.

Obaid, F., Babadi, A., & Yoosofan, A. (2020). Hand Gesture Recognition in Video Sequences Using Deep Convolutional and Recurrent Neural Networks. Applied Computer Systems, 25(1), 57-61. https://doi.org/10.2478/acss-2020-0007

Osman Hashi, A., Zaiton Mohd Hashim, S., & Bte Asamah, A. (2024). A Systematic Review of Hand Gesture Recognition: An Update from 2018 To 2024. IEEE Access, 12, 143599-143626. https://doi.org/10.1109/ACCESS.2024.3421992

Parveen, N., Roy, A., & Sandesh, D. S. (2020). Human-Computer Interaction Through Hand Gesture Recognition Technology. International Journal of Computing and Digital Systems, 9(4).

Rahman, M. M., Uzzaman, A., Khatun, F., Aktaruzzaman, M., & Siddique, N. (2025). A Comparative Study of Advanced Technologies and Methods in Hand Gesture Analysis and Recognition Systems. Expert Systems with Applications, 266(C). https://doi.org/10.1016/j.eswa.2024.125929

Rudwan, M. S. M., & Fonou-Dombeu, J. V. (2023). Hybridizing Fuzzy String Matching and Machine Learning for Improved Ontology Alignment. Future Internet, 15(7), Article 7.https://doi.org/10.3390/fi15070229

Shaikh, M. B., & Chai, D. (2021). Rgb-D Data-Based Action Recognition: A Review. Sensors, 21(12), Article 12. https://doi.org/10.3390/s21124246

Sun, Y., Weng, Y., Luo, B., Li, G., Tao, B., Jiang, D., & Chen, D. (2023). Gesture Recognition Algorithm Based on Multi-Scale Feature Fusion in Rgb-D Images. IET Image Processing, 17(4), 1280-1290. https://doi.org/10.1049/ipr2.12712

Wu, Y., Huang, D., Du, W.-C., Wu, M., & Li, C.-Z. (2020). Joint-Based Hand Gesture Recognition Using RealSense. Journal of Computer, 31(2), 141-151. https://doi.org/10.3966/199115992020043102013

Yasen, M., & Jusoh, S. (2019). A Systematic Review on Hand Gesture Recognition Techniques, Challenges, and Applications. PeerJ Computer Science, 5, e218. https://doi.org/10.7717/peerj-cs.218