Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.071850

ARTICLE

Novel Quantum-Integrated CNN Model for Improved Human Activity Recognition in Smart Surveillance

Tanvir Fatima Naik Bukht[1,2], Yanfeng Wu[1], Nouf Abdullah Almujally[3], Shuoa S. AItarbi[4], Hameedur Rahman[2], Ahmad Jalal[2,5,*] and Hui Liu[1,6,7,*]

[1]Guodian Nanjing Automation Co., Ltd., Nanjing, 210032, China
[2]Department of Computer Science, Air University, E-9, Islamabad, 44000, Pakistan
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11671, Saudi Arabia
[4]Department of Cyber Security, College of Humanities, Umm Al-Qura University, Mecca, 24381, Saudi Arabia
[5]Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul, 02841, Republic of Korea
[6]School of Artificial Intelligence/School of Future Technology, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[7]Cognitive Systems Lab, University of Bremen, Bremen, 28359, Germany

*Corresponding Authors: Ahmad Jalal. Email: ahmadjalal@mail.au.edu.pk; Hui Liu. Email: hui.liu@uni-bremen.de

Received: 13 August 2025; Accepted: 06 November 2025; Published: 23 December 2025

ABSTRACT: Human activity recognition (HAR) is crucial in fields like robotics, surveillance, and healthcare, enabling systems to understand and respond to human actions. Current models often struggle with complex datasets, making accurate recognition challenging. This study proposes a quantum-integrated Convolutional Neural Network (QI-CNN) to enhance HAR performance. The traditional models demonstrate weak performance in transferring learned knowledge between diverse complex data collections, including D3D-HOI and Sysu 3D HOI. HAR requires better extraction models and techniques that must address current challenges to achieve improved accuracy and scalability. The model aims to enhance HAR task performance by combining quantum computing components with classical CNN approaches. The framework begins with bilateral filter (BF) enhancement of images and then implements multi-object tracking (MOT) in conjunction with felzenszwalb superpixel segmentation for object detection and segmentation. The watershed algorithm refines the united superpixels to create more accurate object boundary definitions. The model combination of histogram of oriented gradients (HoG) and Global Image Semantic Texture (GIST) descriptors alongside a new approach to extract 23-joint keypoints by employing relative joint angles and joint proximity measures. A fuzzy optimization process optimizes features that originated from the extraction phase. Our approach achieves 93.02% accuracy on the D3D-HOI dataset and 97.38% on the Sysu 3D HOI dataset Our approach achieves 93.02% accuracy on the D3D-HOI dataset and 97.38% on the Sysu 3D HOI dataset. Averaging across all classes, the proposed model yields 93.3% precision, 92.6% recall, 92.3% F1-score, 89.1% specificity, an False Positive Rate (FPR) of 10.9% and a mean log-loss of 0.134 on the D3D-HOI dataset, while on the Sysu 3D HOI dataset the corresponding values are 98.4% precision, 98.6% recall, 98.4% F1-score, 99.0% specificity, 1.0% FPR and a log-loss of 0.058. These results demonstrate that the quantum integrated CNN significantly improves feature extraction and model optimisation.

KEYWORDS: Quantum-Integrated CNN; image segmentation; computer vision; object detection

1 Introduction

HAR is a critical component in computer vision, enabling systems to understand human actions in various contexts, such as surveillance, robotics, and healthcare [1]. HAR's applications are vast, ranging from improving the accuracy of security surveillance [2] to enabling robots to interact effectively with their environment. However, recognizing complex human interactions remains a challenging task for existing methods [3], which often struggle to generalize across diverse everyday activities, especially when human-object relationships are dynamic and occluded. The primary issue with traditional CNN [4] models is their inability to fully capture temporal relationships and adapt to the extensive variations in human activities across different datasets. This research introduces a novel QI-CNN that combines quantum computing with traditional CNN techniques, aiming to enhance the feature extraction process and improve the scalability and accuracy of HAR in complex environments. This study contributes to addressing gaps in current HAR models by exploring the integration of quantum computing with CNNs for feature optimization and learning in complex human interaction scenarios.

We propose a novel QI-CNN for the HAR system to address these limitations. By leveraging quantum computing within a classical CNN framework, our approach aims to enhance feature extraction and optimization, ultimately improving HAR accuracy. The implementations of machine learning models benefit from quantum computing by enabling faster computation of specific operations. The integration of quantum methods into CNNs shows potential for creating more resilient solutions that address large and complex datasets. The primary objective of this work is to develop and evaluate a QI-CNN that combines the strengths of quantum and classical methods to improve HAR performance. The main contributions of this work are as follows:

- We propose a novel quantum-integrated convolutional neural network (QI-CNN) in which classical features are encoded into an 8-qubit quantum state and processed by a three-layer parameterized quantum circuit comprising RX (Rotation-X) gate, RY (Rotation-Y) gate, RZ (Rotation-Z) gate, CNOT (Controlled-NOT) gate, CZ (Controlled-Z) gate. This tightly coupled quantum layer enriches feature mapping beyond what classical CNNs can achieve and differs from prior hybrid approaches that only append shallow quantum modules.
- We propose a multi modal feature extraction strategy combining Histogram of Oriented Gradients (HoG), GIST descriptors, 23 joint skeleton based features and relative joint angles. These features are processed through a fuzzy optimisation layer that adaptively weights each descriptor, improving robustness to occlusions and pose variations.
- We propose a novel quantum-integrated convolutional neural network (QI-CNN) in which classical features are encoded into an 8-qubit quantum state and processed by a three-layer parameterized quantum circuit comprising RX, RY, RZ, CNOT and CZ gates. This tightly coupled quantum layer enriches feature mapping beyond what classical CNNs can achieve and differs from prior hybrid approaches that only append shallow quantum modules.

In our model, the use of quantum computing to optimize features in human interaction recognition, which is a task that requires complex spatial-temporal relationships in complex datasets, has not been comprehensively discussed. This paper combines quantum feature encoding, fuzzy optimization, and CNN-based feature extraction to suggest a new framework that enhances the robustness of activity recognition in complex situations, including those with occlusions and different poses. By combining quantum computing techniques with classical CNNs and fuzzy optimisation, this work opens a new direction for human activity recognition. The proposed model provides a blueprint for integrating quantum layers into existing deep learning architectures and demonstrates measurable performance gains.

The remainder of the paper is organized as follows: Section 2 reviews related work in HAR, highlighting the limitations of existing methods. Section 3 describes our methodology, including the QI-CNN architecture, feature extraction process, and optimization techniques. Section 4 presents the experimental setup and results, comparing the QI-CNN with traditional CNN models, and a discussion in Section 5. Finally, Section 6 concludes the paper and discusses future directions for quantum-assisted computer vision.

2 Literature Review

The field of computer vision requires HAR because systems need to observe how humans interact with objects. HAR primarily relies on deep learning techniques, particularly through CNNs, as these networks yield excellent results for image classification and object detection tasks. For detecting complicated human-object interactions, machines need to locate objects while simultaneously understanding human body postures and their partnership with items. Meedinti et al. [5] examine the comparative performance of Quantum Convolutional Neural Networks (QCNNs) against traditional CNNs and Artificial Neural Networks (ANNs) in the domains of object detection and classification. Develop a QCNN that utilizes qubits and is architecturally analogous to a CNN, and subsequently evaluate its performance across various models. This study underscores the potential of QCNNs to surpass the performance and efficiency of classical CNNs and ANNs in specific tasks, indicating their promise for real-time image analysis applications. Presently, the limitations of current quantum computing technology preclude experimental investigations on large datasets, necessitating the extension of our knowledge through available means. Liu et al. (2021) [6] suggest combining both quantum and classical convolutional layers to improve the way the model extracts features. Design a quantum convolutional layer using parameterized quantum circuits that feed into classical pooling and fully connected layers; implement an automatic gradient framework; and validate on a Tetris dataset. The hybrid model (QCCNN) achieved higher learning accuracy compared to classical CNN with a similar architecture. The reduced qubit requirements make it suitable for noisy intermediate-scale quantum (NISQ) devices. Current quantum constraints limit circuit depth and qubit numbers; further tests on more complex and larger-scale datasets are required to generalize the findings. The initial limitations of machine learning models, due to feature extraction needs and resource constraints, enabled the recent development of more powerful and robust methods. Traditional feature engineering now pairs with more complex models in multi-stage approaches.

The research shows CNNs achieve high performance in visual recognition but fail to recognize human-object interaction tasks effectively. The capability of classical CNNs to generate effective generalizations declines as these networks handle various datasets with their complex and dynamic human-object relations. These models face a primary challenge because they cannot detect advanced relational characteristics that describe object handling by people in contexts of human-object connection. Ranga et al. (2024) [7] review and analyze various classical-to-quantum data encoding methods in quantum machine learning (QML) and their impact on performance. Survey and compare common encoding strategies (basis, amplitude, angle, etc.) by examining their computational complexity, noise sensitivity, and scalability, and synthesize the literature trends and challenges. Provides a comprehensive comparison of encoding techniques, highlighting how the choice of encoding can impact algorithm efficiency and accuracy. It outlines significant challenges, including scalability, computational burden, and noise issues, and recommends future research directions. Ling et al. (2024) [8] investigate the integration of quantum circuits with classical CNNs for feature extraction and learning efficiency and propose an automatic differentiation framework for training hybrid quantum-classical convolutional networks to implement backpropagation in quantum circuits. Quantum circuits can be embedded efficiently in classical CNN architectures, improving convergence speed and reducing computational overhead. We overview the techniques, findings, weaknesses, and strengths in Table 1.

Table 1: Overview of techniques, key findings, strengths, and weaknesses in HAR detection models

Author(s) & Year	Method	Key findings	Strengths	Weaknesses
Ren et al. (2020) [9]	Rank Pooling CNN segmented bidirectional	Proposed a method that fuses multi-modal RGB and depth data for human action recognition, achieving SOTA performance on multiple datasets.	Effective fusion of RGB and depth data leading to high recognition accuracy.	Performance drops in cases of occlusion or motion in RGB-D sequences.
Hou et al. (2020) [10]	Visual compositional learning	Proposed a method that learns compositional structures for more accurate interaction detection. Achieved higher accuracy in recognizing interactions.	Improves recognition accuracy through compositional learning, particularly in complex interactions.	Struggles with identifying interactions where object categories overlap or are ambiguous.
Zhang et al. (2021) [11]	SCG	Proposed SCG, a graph-based method for detecting human-object interactions, achieves good results on the HICO-DET dataset.	High performance on HICO-DET dataset using graph-based methods for HOI recognition.	Model performance suffers without high-quality human pose estimation.
Kim et al. (2023) [12]	Embedding network multiplex relation	Proposed a three-branch architecture utilizing multiplex relation context for relational reasoning, achieving SOTA performance on HICO-DET and V-COCO.	Effective relational reasoning in multi-modal contexts improves HOI detection accuracy.	Limited context exchange between branches, reducing potential accuracy gains.
Park et al. (2023) [13]	ViPLO	Proposed ViPLO, a two-stage model using ViT backbone and an MOA module for improved feature extraction and interaction detection. Achieved a +2.07 mAP improvement.	Achieved significant improvement (+2.07 mAP) in HOI detection with a ViT backbone.	Requires large computational resources for training the Vision Transformer model.
Quan et al. (2024) [14]	Pairwise CNN-transformer	Developed a two-stage model combining CNN and Transformer features for better HOI detection, improving detection accuracy.	High accuracy in HOI detection using a combination of CNN and Transformer features.	Performance is hindered by occlusion and misclassification of objects in certain scenarios.
Li R et al. (2024) [15]	Bi-FPN and YOLOv4 with ASPP	Enhanced YOLOv4 (ABYOLOv4) by integrating ASPP and Bi-FPN for multi-scale feature fusion, improving accuracy and reducing model size. Achieved a 0.5% AP improvement.	Improved AP and reduced model size, leading to more efficient object detection.	Struggles with occlusion and small targets in complex backgrounds.
Gavali and Kakarwal (2025) [16]	3D-CNN	Proposed a method that decomposes actions into local sub-actions for better early prediction, achieving 82.5% accuracy on SYSU 3D HOI.	Achieved notable early prediction accuracy by focusing on local actions and observing fewer frames.	Challenges in handling visually similar actions and issues with temporal decomposition.

Rank Pooling CNN [9] and Pairwise CNN-Transformer [14] have issues with occlusion and motion, which tend to cause performance deterioration in dynamic settings. Moreover, methods like Visual Compositional Learning [10] have difficulties with ambiguous or overlapping categories of objects, thereby restricting the precision of interaction recognition in complicated situations. Other models, such as Spatially Conditioned Graph (SCG) [11] are also sensitive to the quality of human pose estimation and therefore they are susceptible to performance degradation when the quality of pose data is compromised. Also, other approaches like ViPLO [13] and 3D-CNN [16] are computationally intensive and thus challenging to implement in resource-limited settings. The other gap is seen in the Embedding Network with Multiplex Relation [12], where the context exchange between the branches is limited, which prevents the model from working well on complex interactions. Finally, models such as 3D-CNN [16] have problems distinguishing between visually similar actions, which results in recognition problems in case the actions are visually similar but different in context.

Our proposed QI-CNN model addresses these gaps in several ways. By integrating quantum-enhanced feature extraction, our model is significantly more robust to occlusion and motion compared to previous methods, leading to superior performance in dynamic and cluttered environments. The quantum layer enables our model to deal with overlapping or ambiguous object categories more effectively, enhancing the accuracy of interaction detection. In contrast to SCG [11], our model is not based on pose estimation, which is more practical in real-life situations where pose data is not perfect or absent. Computationally, our hybrid quantum-classical model is optimal in both performance and computational cost, and is scalable even in resource-constrained environments. The quantum optimization also improves contextual reasoning among model components, which is limited in context exchange in previous models, such as Multiplex Relation Embedding Network [12]. Lastly, the quantum layer enhances the model to differentiate visually similar actions, providing a superior differentiation in action recognition tasks, compared to 3D-CNN [16]. Our QI-CNN model offers a robust, scalable, and computationally efficient solution to the problems of human-object interaction recognition through these innovations.

3 Materials and Methods

3.1 System Methodology

The proposed approach for recognizing HAR combines traditional CNNs with quantum computing to improve both accuracy and efficiency. The process begins by selecting the D3D-HOI and Sysu 3D HOI datasets, which feature various HAR scenarios. In the preprocessing stage, a bilateral filter is used to clean the data by reducing noise while keeping important image details intact. For human detection and segmentation, we use MOT to follow objects within the frame, along with the Felzenszwalb superpixel segmentation algorithm to group pixels based on similar textures and colors. Feature extraction is done in two steps: skeleton features identify 23 key joint points for human pose analysis, and visual features combine HoG and GIST descriptors to recognize background and objects. These features are then optimized through fuzzy techniques to improve their relevance. At the heart of this approach is a hybrid model that integrates classical algorithms with quantum computing. This combination enhances object detection and feature segmentation, leading to better overall performance as shown in Fig. 1.

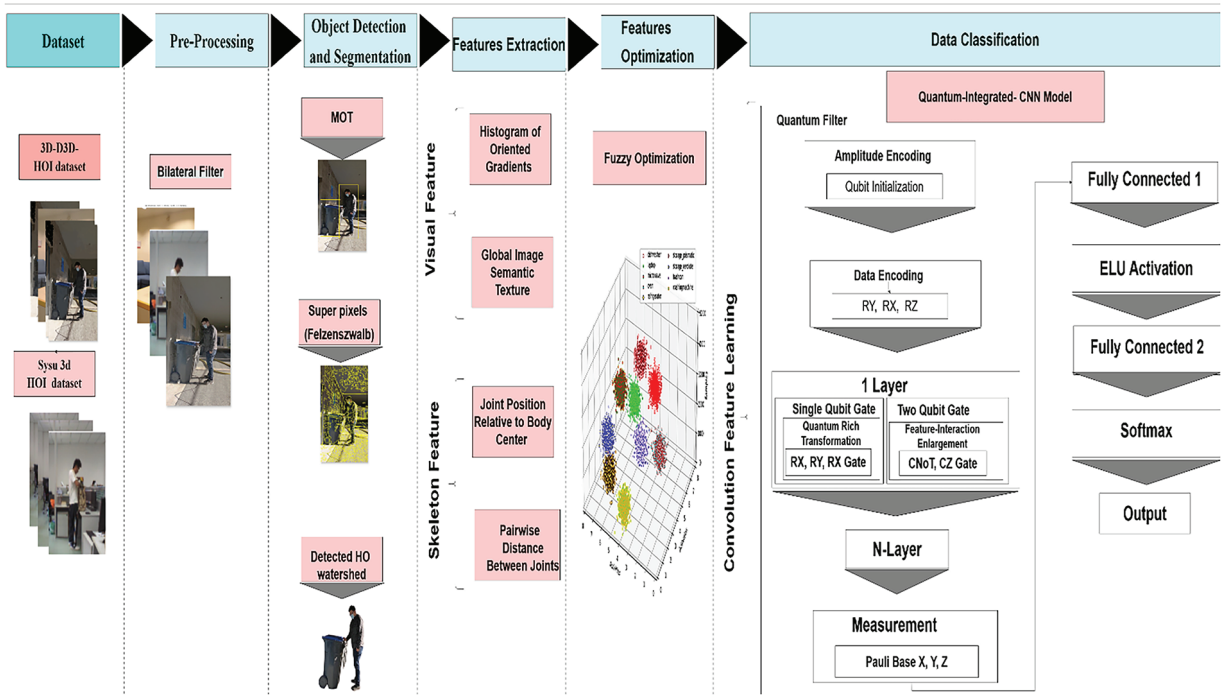


Figure 1: The proposed methodology illustrates the integration of quantum computing with CNNs for enhanced HAR

BF helps in preserving important edges while reducing noise in the images, which is essential for improving object detection and feature extraction. MOT ensures consistent object identities across frames, a crucial step when working with dynamic datasets, reducing the risk of errors due to misidentified objects. Superpixel Segmentation and Watershed Algorithm enhance the segmentation process, providing more accurate boundaries for detected objects, particularly in complex scenes with occlusions. Fuzzy Optimization is employed to improve feature selection, optimizing the performance of the model by handling uncertain and ambiguous features more effectively than traditional methods like Principal Component Analysis (PCA). The Quantum Layer enables the model to learn complex feature transformations that are difficult to achieve with classical approaches alone, making the process more robust in dealing with complex, non-linear relationships in the data.

3.2 Pre-Processing

It is essential to enhance the original images by preprocessing them before they are utilized in subsequent stages. In this paper, a BF is applied to remove noise while preserving the edges of images. BF is a type of nonlinear image filtering that smoothes areas by considering the proximity of pixels to each other and the difference in their brightness. With this filter, blurry areas are cleaned, and spots with noticeable shifts in lighting are left unchanged. The BF consists of a weight function and an average calculation that utilizes these weights.

$$\boldsymbol{w}(p,q) = \exp\left(-\frac{|p-q|^2}{2\sigma_d^2}\right)\exp\left(-\frac{|I(p)-I(q)|^2}{2\sigma_r^2}\right) \tag{1}$$

The weight function Eq. (1) is $(\boldsymbol{w}(p,q))$ merges spatial distance with pixel intensity similarity between the central pixel $(\boldsymbol{p})$ and neighboring pixel (q). The spatial difference between pixels is denoted by $(|\boldsymbol{p}-\boldsymbol{q}|)$,

and the intensity difference is represented by $(|\boldsymbol{I}(p) - \boldsymbol{I}(q)|)$. The parameters control the spatial and intensity domain smoothing amounts (σ_d) and (σ_r). This research adopted an empirical approach to select the parameter σ_d at 5 to control spatial smoothing and maintain edges. A value of 30 was assigned to the intensity parameter σ_r to achieve optimal edge preservation while clearing noise from homogeneous areas. Lower σ_d settings would have maintained edges too well yet failed to eliminate sufficient noise from flat areas. Higher values of σ_r produce smoothing effects result in edge blurring, thus making the technique inappropriate for maintaining edge sharpness. The selected values demonstrated an appropriate trade-off for image processing in this research project. BF considers two major factors. The first step relies on the Gaussian function to evaluate the weight depending on the distance between different pixels. Next, assign a weight to each pixel by comparing its intensity to that of its surrounding pixels. Since pixel intensity determines how the weight is distributed, sharp edges are preserved.

3.3 Object Detection and Seqmentation

Predicting HAR requires systems that detect targets with precision, as this ability directly determines the pathway to separate human beings from their associated objects. Multiple stages organize the process for segment refinement along with HAR tracking. The four essential steps include MOT, superpixel segmentation using felzenszwalb, merging of superpixels, and human-object watershed segmentation.

3.3.1 Multi-Object Tracking (MOT)

The object detection pipeline begins with MOT as its initial process. MOT serves to detect and monitor multiple objects, including humans and other objects, through consecutive video frames. The detection process creates moving boxes that connect to identified objects throughout time for their precise localization and tracking. Binding boxes function as regions of interest to support the subsequent processing stage.

3.3.2 Superpixels (Felzenszwalb)

After object tracking using the felzenszwalb algorithm, image segmentation using superpixels. By grouping pixels into superpixels, the felzenszwalb algorithm exhibits common characteristics like color, texture, or intensity levels. After segmentation, the next step in watershed detection is to determine. Felzenszwalb's method outperforms the boundary-driven and quick SLIC (Simple Linear Iterative Clustering) technique for handling detailed images. Because the watershed algorithm accurately identifies objects that overlap, it is better suited for studying how people interact with objects. Complex visual scenes can be analyzed more effectively when watershed detection is combined with Felzenszwalb's superpixel segmentation [17]. The Felzenszwalb algorithm combines nearby pixels with similar properties. First, the image is broken up into many small parts, which are then combined because they are all the same size and intensity. Each superpixel in the image indicates an area with a similar set of pixels. When determining whether to combine two clusters, a similar function S is defined as follows.

$$\boldsymbol{S}(\boldsymbol{A},\boldsymbol{B}) = \frac{|\boldsymbol{I_A} - \boldsymbol{I_B}|^2}{\boldsymbol{\sigma_d^2}} + \frac{|\boldsymbol{P_A} - \boldsymbol{P_B}|^2}{\boldsymbol{\sigma_r^2}} \tag{2}$$

where $(\boldsymbol{A})$ and $(\boldsymbol{B})$ represent two neighboring superpixels, $(\boldsymbol{I_A})$ and $(\boldsymbol{I_B})$ are the intensity values of the superpixels, $(\boldsymbol{P_A})$ and $(\boldsymbol{P_B})$ are their spatial coordinates, and $(\boldsymbol{\sigma_d})$, $(\boldsymbol{\sigma_r})$ control the influence of spatial and intensity differences. The similarity function enables the merging of superpixels that are spatially close and share similar intensity characteristics, which aids in object segmentation, as shown in Fig. 2.

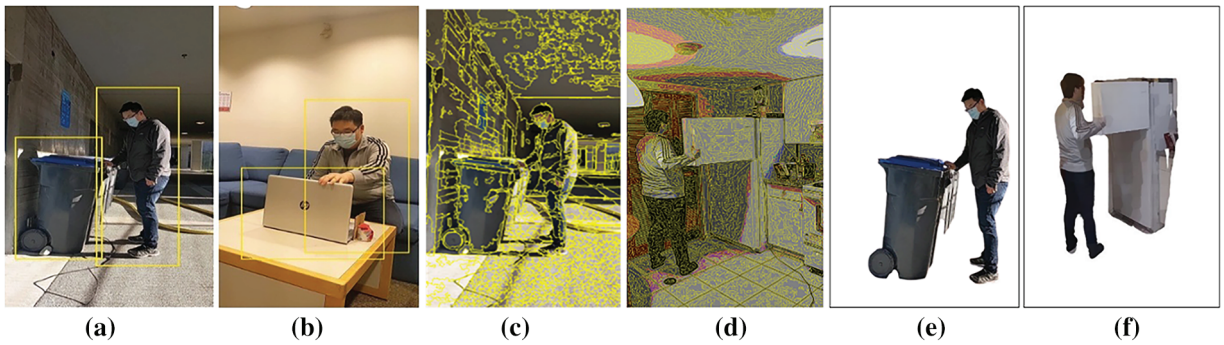


Figure 2: The figure illustrates the results of object detection and segmentation, results of MOT (**a**) trashcan (**b**) laptop, results of Felzenszwalb: (**c**) opening a trashcan, (**d**) opening a refrigerator and the detected human object watershed (**e**) opening a trashcan, (**f**) opening a refrigerator

3.3.3 Detected HO Watershed

The watershed algorithm is based on the concept of image flooding, where the lowest regions (basins) are filled first, and the boundaries are formed at higher elevation levels. The watershed transform $(\boldsymbol{W})$ for an image $(\boldsymbol{I})$ is defined as below.

$$\boldsymbol{W}(\boldsymbol{I}) = \textbf{argmax}\left(\nabla \boldsymbol{I}(\boldsymbol{x}, \boldsymbol{y})\right) \tag{3}$$

where $(\nabla \boldsymbol{I}(\boldsymbol{x}, \boldsymbol{y}))$ represents the gradient of the image intensity at pixel location $(\boldsymbol{x}, \boldsymbol{y})$. The gradient indicates the boundary between different regions, and the watershed algorithm uses these gradients to segment the image into distinct regions.

The results of HAR tracking using MOT are presented in Fig. 3 across multiple video frames. A similar function enables the merging of superpixels that are spatially close and share similar intensity characteristics, which aids in object segmentation, as shown in Fig. 2. The human-object watershed segmentation ensures that the boundaries between humans and objects are accurately detected, even in complex scenes with occlusion or close interactions, as shown in Fig. 2.

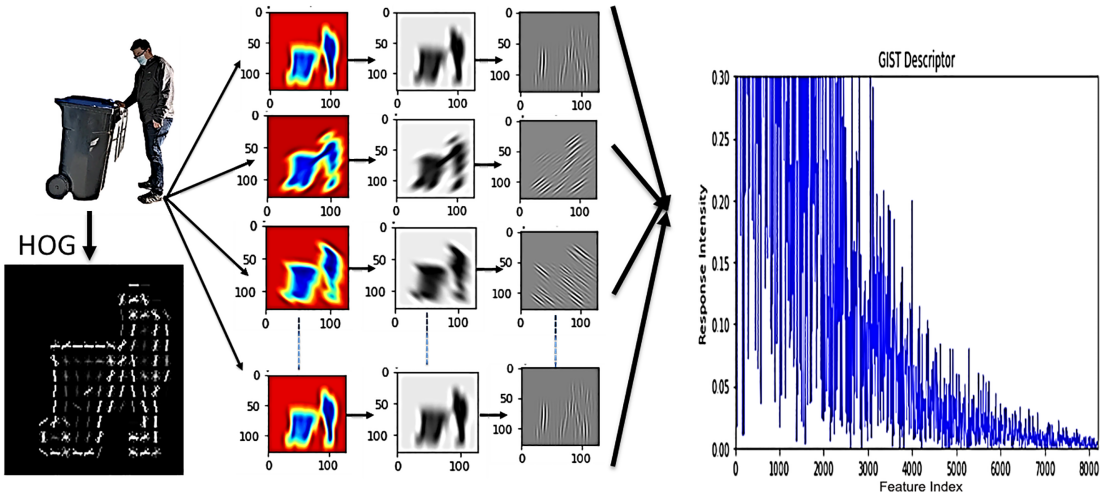


Figure 3: The analysis of GIST feature extraction focuses on global spatial data as a vital factor for detecting HAR in complex visual environments and Results of HOG

3.4 Feature Extraction

The HAR heavily depends on feature extraction; the research focuses on extracting two fundamental features, including visual-specific elements and skeleton-based characteristics.

3.4.1 Visual Feature Extraction

Visual aspects focus on texture, as well as on shape and structure, to assist in analyzing how humans interact with objects in various situations. By examining the gradient values of x and y, the HoG algorithm obtains data on local edges and object outlines. It divides the image into cells of size 8 × 8 and arranges the changes in light and space by combining gradients into only nine bins.

Histogram of Oriented Gradients (HoG)

Computer vision applications heavily rely on HoG as one of their primary feature descriptor methods. Local image regions contain edge and gradient structures that HoG effectively detects for detection.

$$\boldsymbol{H}_{\text{cell}}(\boldsymbol{i}) = \sum_{\boldsymbol{p}\in\text{cell}} \boldsymbol{w}(\boldsymbol{p}) \cdot \delta(\theta(\boldsymbol{p}) - \boldsymbol{i}) \tag{4}$$

In Eq. (4), a histogram for each cell is constructed by assigning each pixel in the cell to one of the orientation bins based on the computed gradient direction. where $\boldsymbol{w}(\boldsymbol{p})$ is the weight (usually the gradient magnitude $\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{y})$) for pixel $\boldsymbol{p}$, $\theta(\boldsymbol{p})$ is the gradient direction at pixel $\boldsymbol{p}$, and $\delta(\theta(\boldsymbol{p}) - \boldsymbol{i})$ is the Dirac delta function that assigns the pixel to the appropriate orientation bin $\boldsymbol{i}$. To account for lighting changes, the histograms for cells within a block are concatenated and normalized. Then, the HoG descriptor for the entire image is the concatenation of the normalized histograms from all blocks. The HoG descriptor captures the gradient distribution in the image and is robust to changes in lighting and small spatial transformations, as shown in Fig. 4.

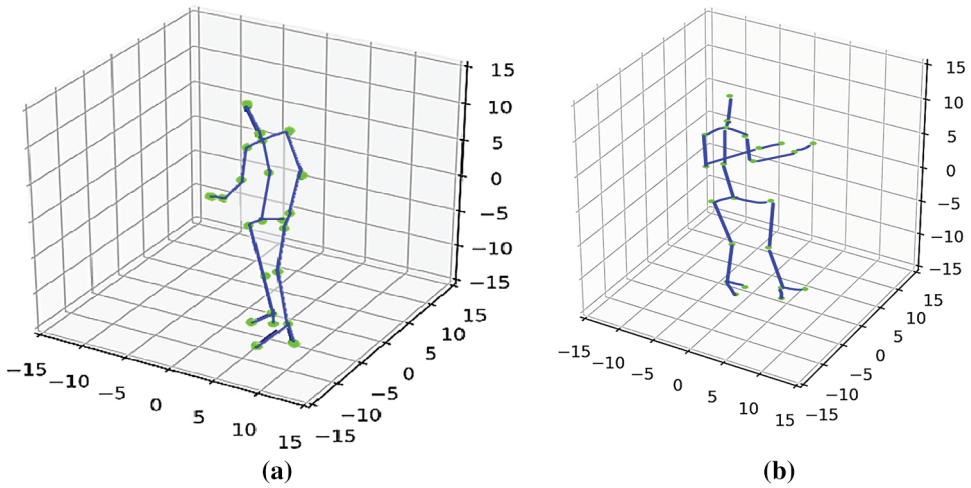


Figure 4: 3D visualization of the skeleton model with the key joints (**a**) opening trashcan (**b**) opening refrigerator

Global Image Semantic Texture (GIST)

GIST is a global feature descriptor that captures a visual's overall spatial layout and texture. It is beneficial for recognizing a visual's global context, which focuses on local details. GIST is obtained by applying a set of spatial filters at multiple scales and orientations, which capture an image's broad spatial and textural properties. GIST utilizes a filter bank comprising a set of oriented filters at various scales. This filter is based

on the Gabor function, which captures both frequency and orientation information.

$$\boldsymbol{R}_{\text{filter}}(\boldsymbol{s},\theta) = \sum_{\boldsymbol{x},\boldsymbol{y}} \boldsymbol{I}(\boldsymbol{x},\boldsymbol{y}) \cdot \psi_{\boldsymbol{s}}(\boldsymbol{x},\boldsymbol{y},\theta) \tag{5}$$

where in Eq. (5) $(\boldsymbol{I}(\boldsymbol{x},\boldsymbol{y}))$ is the intensity of the image at pixel $((\boldsymbol{x},\boldsymbol{y}))$, and $(\psi_{\boldsymbol{s}}(\boldsymbol{x},\boldsymbol{y},\theta))$ is the filter response at scale $(\boldsymbol{s})$ and orientation (θ). To lower the dimensionality, the reactions are grouped into spatial regions after the filters are applied. To accomplish this, the image is divided into small blocks, each measuring 4×4 pixels, and the average response is calculated in each block. Concatenating the responses from each filter at various scales and orientations yields the final GIST descriptor.

The outcomes of GIST feature extraction applied to a picture of a person interacting with a trash can are shown in Fig. 3. The filters allow GIST to create an abstract representation of the entire image by identifying both spatial patterns and image textures.

3.4.2 Skeleton Feature Extraction

The skeletal geometry features are beneficial when the information is obtained from the skeletal structure to determine human motion and behavior. HAR with the novel skeleton model, showing the 23 key joint points (green) and skeleton connections (blue) overlaid on Fig. 4. and 3D representation of the same skeleton model, highlighting the key joint points and the overall structure.

Joint Position Relative to Body Center

We normalize joint positions relative to the body center using the Generalized Procrustes Analysis (GPA), which effectively removes variations caused by translation, scale, and rotation. In Eq. (6) 3D joint coordinates $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_N\}$ of set from a subject's pose, and a reference skeleton $\boldsymbol{Q} = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_N\}$, we also compute the optimal rotation $\boldsymbol{R} \in \boldsymbol{SO}(3)$, scaling factor $\boldsymbol{s}$, and translation vector $(\boldsymbol{t})$ that minimize the sum of squared distances.

$$\min_{\boldsymbol{s},\boldsymbol{R},\boldsymbol{t}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} |\boldsymbol{sRp}_{\boldsymbol{i}} + \boldsymbol{t} - \boldsymbol{q}_{\boldsymbol{i}}|^2 \tag{6}$$

We get joint positions that are normalized as $\widehat{\boldsymbol{p}_{\boldsymbol{i}}} = \boldsymbol{sRp}_{\boldsymbol{i}} + \boldsymbol{t}$. This process centers the skeleton around a canonical body reference, compensates for differences in subject size and orientation, and ensures that pose features are comparable and consistent across diverse datasets. By employing GPA, we enhance the pose representation, which is crucial for accurately recognizing static postures and dynamic activities.

Pairwise Distance between Joints

Calculating the pairwise distances between key joints, which capture the spatial relationship between different parts of the body. We model the human skeleton as a graph $(\boldsymbol{G} = (\boldsymbol{V},\boldsymbol{E}))$, where nodes represent joints and edges represent bones connecting these joints. We calculate the geodesic distance between any two joints $(\boldsymbol{i})$ and $(\boldsymbol{j})$ as the length of the shortest path along the edges of the graph.

$$\boldsymbol{d}_{\boldsymbol{G}}(\boldsymbol{i},\boldsymbol{j}) = \min_{\boldsymbol{P}_{\boldsymbol{i}\to\boldsymbol{j}}} \sum_{(\boldsymbol{u},\boldsymbol{v})\in\boldsymbol{P}_{\boldsymbol{i}\to\boldsymbol{j}}} |\boldsymbol{p}_{\boldsymbol{u}} - \boldsymbol{p}_{\boldsymbol{v}}| \tag{7}$$

Eq. (7) $(\boldsymbol{P}_{\boldsymbol{i}\to\boldsymbol{j}})$ denotes a path on the graph connecting $(\boldsymbol{i})$ and $(\boldsymbol{j})$, and $(|\boldsymbol{p}_{\boldsymbol{u}} - \boldsymbol{p}_{\boldsymbol{v}}|)$ is the Euclidean distance between adjacent joints $(\boldsymbol{u})$ and $(\boldsymbol{v})$. The translation of this geodesic model in terms of body structure provides more significant results for understanding joint relationships, as opposed to dealing with their direct distances in Euclidean space. The distance from the shoulders, elbows, or knees helps describe the exact positioning of the limbs and body parts in a static position.

3.5 Feature Optimization

Optimization of features is a key process in HAR when systems utilize more complex human skeleton joint models. Fuzzy optimization allows decision-makers to resolve the uncertainty or vagueness of data during decision-making.

$$R_i = \sum_{j=1}^{n} \mu\left(F_j\right) \cdot w_j \tag{8}$$

Eq. (8) calculates the feature (F_i) final relevance score (R_i) through multiplication of $(\mu\left(F_j\right))$ membership value and (w_j) weight of feature (F_j). Each feature receives its weight (w_j) which determines its significance in the optimization process. The weights (w_j) allow us to modify feature optimization according to domain-specific needs and empirical data. The final stage of fuzzy optimization involves selecting only the most essential features, which are then subjected to further processing.

In Fig. 5, we present the cluster centroids, cluster sizes, and validation metrics resulting from the fuzzy logic optimization process. The total number of samples used in the clustering is 5000. The centroids of the clusters provide insight into the average feature values within each activity class. For example, Cluster 1 (Dishwasher) has a centroid at (Skeleton: 0.65, 1.2) and (Visual: HoG: 0.75, GIST: 0.88). The size of Cluster 1 is 500 samples. The Silhouette Score for Cluster 1 is 0.85, indicating good cohesion within the cluster and clear separation from others. Additionally, the Davies-Bouldin Index for Cluster 1 is 0.45, suggesting that the clusters are well-separated and compact. These results demonstrate the success of fuzzy logic optimization in effectively grouping similar features while maintaining clear boundaries between different activity classes, even in the presence of uncertainty and overlapping data.

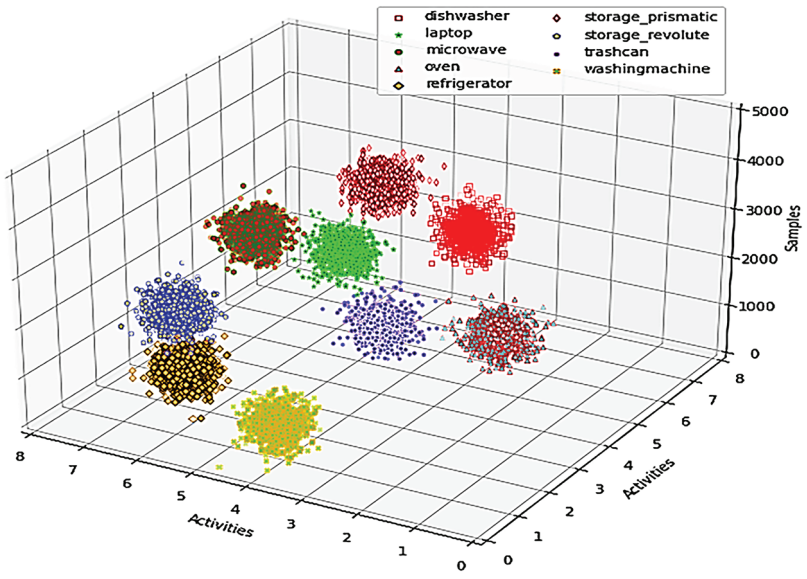


Figure 5: Illustration of Fuzzy Optimization on extracted features of 3D-D3D-HOI

Fuzzy logic surpasses PCA and ICA because it enables human object features and data instances to belong to multiple classes to varying degrees, thus delivering a more accurate representation of natural human actions. The HAR requires special attention to human poses alongside object occlusions and partial interactions since these elements frequently obscure clear class boundaries. Fuzzy logic systems demonstrate robust performance and adaptability in uncertain conditions because they achieve superior results to crisp decision models when processing authoritatively ambiguous or poorly defined data inputs [18]. Zheng, in

his research, demonstrates how treating learning algorithms and fuzzy logic systems together yields better accuracy in gesture and interaction detection, particularly under unreliable conditions [19].

Table 2 are the quantitative interpretation of clusters in Fig. 5, which are the various patterns of purpose usage based on fuzzy optimization of interaction data. The centroid of each cluster represents the average location of the usage pattern of the classes on the three features: Activities, Samples, and Categories. The cluster size is the number of samples in each category, and the cluster spread, which is determined by standard deviation, is the variability in the use of classes. As an example, the refrigerator cluster is more tightly spread (0.4) than the washing machine (1.0), which means that it is used more regularly. The average distance between centroids measures the distinctiveness of the usage pattern of each class relative to others. As an illustration, the laptop and microwave clusters are separated by 1.5 units, with more pronounced usage than the oven and dishwasher, which are separated by 1.2 units. These quantitative measures give a better insight into the patterns of interaction, with more variability in classes (e.g., washing machine) and less variability in classes (e.g., refrigerator). The error margins indicate the accuracy of the cluster assignment, and all classes have comparatively low error margins (between 2.8 and 3.4 percent), which proves the accuracy of the fuzzy optimization process in clustering similar usage patterns.

Table 2: Quantitative interpretation of clusters in Fig. 5

Classes	Cluster centroid (x, y, z)	Cluster size (Number of samples)	Cluster spread (Standard deviation)	Mean distance between centroids	Error margin (%)
Dishwasher	(3.2, 1.1, 5.0)	450	0.5	1.2	2.8%
Laptop	(6.0, 2.3, 4.5)	400	0.7	1.5	3.1%
Microwave	(4.5, 1.8, 3.5)	420	0.6	1.3	2.9%
Oven	(2.9, 1.4, 2.7)	380	0.8	1.0	3.4%
Refrigerator	(3.5, 1.2, 5.5)	470	0.4	1.1	2.7%
Washing machine	(7.0, 2.8, 6.0)	500	1.0	1.6	3.0%
Trashcan	(1.8, 0.9, 4.2)	300	0.5	1.4	3.3%
Storage	(5.5, 2.0, 7.0)	460	0.6	1.2	2.8%

Table 3 shows fuzzy optimization is more complex in time (O(n*m*p)) and time (1.10 s) than PCA (0.75 s) and LDA (0.85 s), its iterative nature enables it to deal with non-linear relationships and uncertainty in data. This leads to reduced error rates (2.9%), which is better in complex data, particularly in human interaction recognition. Conversely, PCA and LDA are good in linear problems but not in situations where data contains noise or non-linearities, which increases the error margins. Although the energy consumption is slightly increased (6.0 J to fuzzy optimization and 4.2 J to PCA), the accuracy and robustness gains are worth the additional computational cost, and fuzzy optimization is the best option for my model.

3.6 Quantum-Integrated Convolutional Neural Network Model

The QI-CNN model is a novel approach to HAR, effectively integrating the strengths of both quantum computing and classical deep learning techniques. This hybrid architecture leverages quantum computing's ability to process high-dimensional data and classical CNN's powerful feature learning and classification capabilities. The model operates in two primary phases: a quantum phase for feature transformation and a classical phase for feature learning and activity classification. The quantum-integrated convolutional neural network model is shown in Fig. 6.

Table 3: Comparative analysis of fuzzy optimization with PCA and LDA **(Linear Discriminant Analysis)**

Method	Time complexity	Execution time (s)	Energy efficiency (J)	Error margins (%)
Fuzzy optimization	O(n*m*p), n = features, m = data points, p = iterations	1.10	6.0	Low (2.9%)
PCA (Principal Component Analysis)	O(n^2), n = features	0.75	4.2	Low (3.0%)
LDA	O(n^2), n = features	0.85	5.0	Moderate (4.0%)

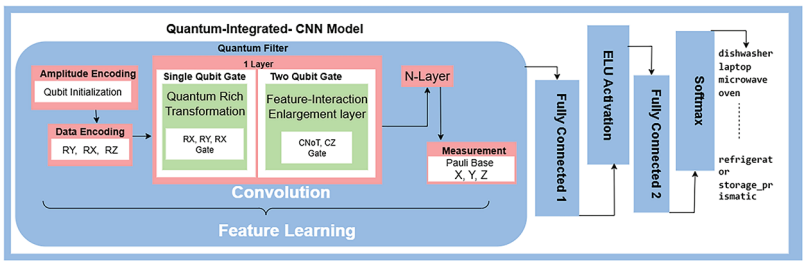


Figure 6: Quantum-integrated convolutional neural network model

The quantum part of the model is simulated on Qiskit Aer's statevector simulator, utilizing an 8-qubit circuit with a three-layer variational ansatz. Each layer consists of single-qubit rotations (RY(θ)) followed by a ring of CNOT gates entangling neighboring qubits. The training of the quantum model is performed by optimizing the variational parameters jointly with the classical CNN weights using the Adam optimiser (learning rate = 0.01, 500 iterations). Each circuit evaluation was repeated 1024 times for accurate measurement, and the state vector simulation was run on a standard laptop with 16 GB RAM.

3.6.1 Quantum Filter and Feature Transformation

The first step in the Novel quantum-integrated CNN Model involves transforming classical feature data into quantum representations. The classic feature vector ($\boldsymbol{F} = [\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_n]^T$) is encoded into a quantum state by amplitude encoding. The encoding transforms the feature vector by its Euclidean norm and maps it to the states ($\{|\boldsymbol{i}\rangle\}$) in a quantum system in Eq. (9), ending with the quantum state ($|\psi\rangle$). Now, classical data can be directly embedded into the quantum Hilbert space for further processing.

$$|\psi\rangle = \frac{1}{|\boldsymbol{F}|_2} \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{F}_{\boldsymbol{i}} |\boldsymbol{i}\rangle, \text{ where } |\boldsymbol{F}|_2 = \sqrt{\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} |\boldsymbol{F}_{\boldsymbol{i}}|^2} \tag{9}$$

3.6.2 Quantum Gate Operation

The quantum states containing encoded classical features must be manipulated using quantum gates to execute feature enhancement operations. These operations enable the quantum system to detect sophisticated, high-dimensional features, resulting in improved representation. Our quantum system applies RX,

RY, and RZ gates to individual qubits, modifying their states. A quantum transformation occurs through these gates. Then, several parametrized quantum gate layers are applied to the encoded state by the quantum filter. Every layer only uses rotation modules to turn one qubit at a time, then connects them with entangling operations. The combined unitary transformation $\boldsymbol{U}(\theta)$ is built by stacking $\boldsymbol{L}$ layers. After this, the result for the state $|\psi\rangle$ is a new state $|\psi_{\text{out}}\rangle$ that highlights the interaction between various quantum features.

$$\boldsymbol{s}\,|\psi_{\text{out}}\rangle = \underbrace{\prod_{\boldsymbol{l}=1}^{\boldsymbol{L}}\left(\bigotimes_{\boldsymbol{k}=1}^{\boldsymbol{m}}\boldsymbol{R_k}\left(\boldsymbol{\theta_l}\right)\right)\cdot\boldsymbol{E_l}}_{\boldsymbol{U}(\theta)}|\psi\rangle\,, \tag{10}$$

In Eq. (10), $(\boldsymbol{R_k}(\theta_l))$ denotes the operator for a single-qubit rotation on the $(\boldsymbol{k}^{\text{th}})$ qubit at layer $(\boldsymbol{l})$ parameterized by (θ_l) and $(\boldsymbol{E_l})$ represents how qubits are entangled at that layer. Using these layers, the model can learn essential features that aid traditional tasks. $\bigotimes_{\boldsymbol{k}=1}^{\boldsymbol{m}}\boldsymbol{R_k}\left(\boldsymbol{\theta_l}\right)$ represents an independent single-qubit rotation gate. Two-qubit gates, such as the CNOT (Controlled-NOT) and CZ (Controlled-Z) gates, are used to introduce quantum entanglement between qubits. The use of quantum operations is effective in mixing the entangled qubits to maximize the feature interaction. The quantum state is then measured with the help of the Pauli basis (X, Y, Z) and then it is collapsed to the classical data and further processed by CCNN.

3.6.3 Classical CNN for Feature Learning and Classification

The classical CNN is responsible for learning the hierarchical patterns within the transformed features and performing the final classification task. After the quantum measurement, the classical features F quantum is passed to the first fully connected layer F_{fc1} = W1 · F quantum + b1 where W1 is the weight matrix, b1 is the bias term, and F quantum is the result from the quantum measurement. After the fully connected layer 1, the output is passed through the Exponential Linear Unit (ELU) activation function, which introduces non-linearity into the model. The ELU activation function is defined as:

$$F_{elu}(x,y) = \begin{cases} F_{fc1}(x,y) & if\ F_{fc1}(x,y) > 0 \\ \alpha\left(e^{F_{fc1}(x,y)} - 1\right) & if\ F_{fc1}(x,y) \le 0 \end{cases} \tag{11}$$

In Eq. (11), (α) is typically set to 1, $F_{fc1}(\boldsymbol{x},\boldsymbol{y})$ is the output from the first fully connected layer. The output of the ELU activation is fed into the second fully connected layer, F_{fc2} = W2 Felu + b2, where W2 is the weight matrix and b2 is the bias. The last layer is the output layer that applies the softmax activation function to transform the scores of classes into probabilities. The QI-CNN model integrates classical CNNs to learn features and classify them with quantum computing to transform features. This framework improves HAR through the integration of quantum gates and classical CNNs to generate improved feature spaces to track activities. Classical CNN is used to classify activities whereas the quantum part is used to determine the complex feature relationships.

Algorithm 1 shows that a hybrid system is the combination of quantum computing and traditional CNNs to improve human activity recognition. The quantum feature transformation makes the training quicker and allows recognition in an instant. The model utilizes quantum computing to combine CNNs with feature transformation to a higher dimension to enhance the classification of the activities.

Algorithm 1: Novel quantum-integrated CNN Model for Human interaction Recognition

1: Input: Feature
2: Output: Predicted human interaction class yclass
3: Quantum Feature Transformation
4: Encode classical features F into quantum state $|\psi\rangle$ using amplitude encoding:

$$|\psi\rangle = \frac{1}{|F|_2} \sum_{i=1}^{n} F_i \, |i\rangle ,$$

where $|\mathrm{F}|_2 = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{F_i}|^2}$ for each classical feature Fi.
5: Apply quantum gates to modify quantum states: for each quantum feature do
6: Apply single-qubit gates (RX, RY, RZ) for quantum feature transformation.
7: Apply two-qubit gates (CNOT, CZ) for entanglement and feature interaction.
8: Apply Feature-Interaction Enlargement Layer to further enhance feature interactions.
9: Measure quantum state using Pauli basis (X, Y, Z) to extract classical data:

$$Fquantum = Measurement \; (|\psi enhanced\rangle)$$

10: Classical CNN Processing
11: Pass Fquantum into the classical CNN for classification.
12: Apply fully connected layer 1: F_{fc1} = W1 · Fquantum + b1
13: Apply ELU activation:

$$F_{elu}\,(x, y) = \begin{cases} F_{fc1}\,(x, y) \\ \alpha \left(e^{F_{fc1}(x,y)} - 1 \right) \end{cases}$$

14: Apply fully connected layer 2: F_{fc2} = W2 · Felu + b2
15: Apply softmax activation to get class probabilities: y class = Softmax(F_{fc2})
16: Return y_class⟵predicted human-interaction class

4 Dataset Experimental Setup and Results

The experiments were conducted on a laptop equipped with a Windows 11 operating system, an AMD Ryzen 7000 series processor with Radeon Graphics, and a 2.00 GHz speed.

4.1 Dataset Description

The D3D-HOI dataset comprises videos with 3D object pose, shape, and part motion during human interactions. The dataset comprises 256 videos, distributed across eight categories, and provides 3D labels for object rotation and translation, as well as size and part motion information for each frame [20]. This dataset comprises everyday articulated objects captured from various real-world visuals and viewpoints, representing each manipulated object through matching 3D parametric models.

The SYSU 3D HOI dataset comprises 480 video clips that cover 12 different activity classes. The dataset includes interactions between 40 subjects who used RGB-D sensors to perform actions with various objects, including phones, chairs, bags, wallets, mops, and brooms. The dataset comprises 40 samples per activity, making it suitable for activity recognition studies [21] in various environmental settings.

4.2 Result Evaluation

To mitigate the risk of overfitting, we employed 10-fold cross-validation, which helps to ensure that the model is evaluated on multiple subsets of the data, maximizing the use of available samples and providing a more reliable estimate of performance. This technique helps in reducing variance and ensures that the model's performance is not overly dependent on any single data split. Additionally, we applied regularization

techniques in the training process to further prevent overfitting, ensuring that the proposed pipeline remains robust despite the limited dataset size. We believe that these steps, along with the rigorous evaluation provided by cross-validation, address concerns related to overfitting and enhance the generalizability of the results. The performance evaluation of the novel quantum-integrated CNN model focuses. The precision (Pr), recall (Rc), Specificity (Sp), Log Loss (LL)and F1-score (F1) for each class are summarized in Table 4 and confusion matrix in Fig. 7. Eq. (12) accuracy is the proportion of correctly predicted instances, both true positives and true negatives, to the total number of instances where: TrPr = True Positives, TrNe = True Negatives, FaPo = False Positives, FaNe = False Negatives. Eq. (13) precision indicates the proportion of true positive results among all optimistic predictions made by the model. Eq. (14) recall, or sensitivity, measures the proportion of actual positives that the model correctly identified. Eq. (15) F1-score is the harmonic means of precision and recall, providing a single measure that balances both metrics. Eq. (16) specificity is the proportion of true negatives identified correctly by the model. Eq. (17) False Positive Rate indicates the proportion of negatives that were incorrectly identified as positives. Log Loss, or Logistic Loss, measures the uncertainty of the model's predictions based on how close the predicted probabilities are to the actual labels. In Eq. (18), (y_i) is the true label (0 or 1) of the i-th sample, (p_i) is the predicted probability of the positive class for the i-th sample, (N) is the total number of samples.

$$\text{Accuracy} = \frac{\text{TrPr} + \text{TrNe}}{\text{TrPr} + \text{TrNe} + \text{FaPo} + \text{FaNe}} \tag{12}$$

$$\text{Precision} = \frac{\text{TrPr}}{\text{TrPr} + \text{FaPo}} \tag{13}$$

$$\text{Recall} = \frac{\text{TrPr}}{\text{TrPr} + \text{FaNe}} \tag{14}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

$$\text{Specificity} = \frac{\text{TrNe}}{\text{TrNr} + \text{FaPo}} \tag{16}$$

$$\text{FPR} = \frac{\text{FaPo}}{\text{FaPo} + \text{TrNe}} \tag{17}$$

$$\text{LL} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \tag{18}$$

Table 4: Detailed performance metrics on the 3D-D3D-HOI and Sysu 3D HOI dataset

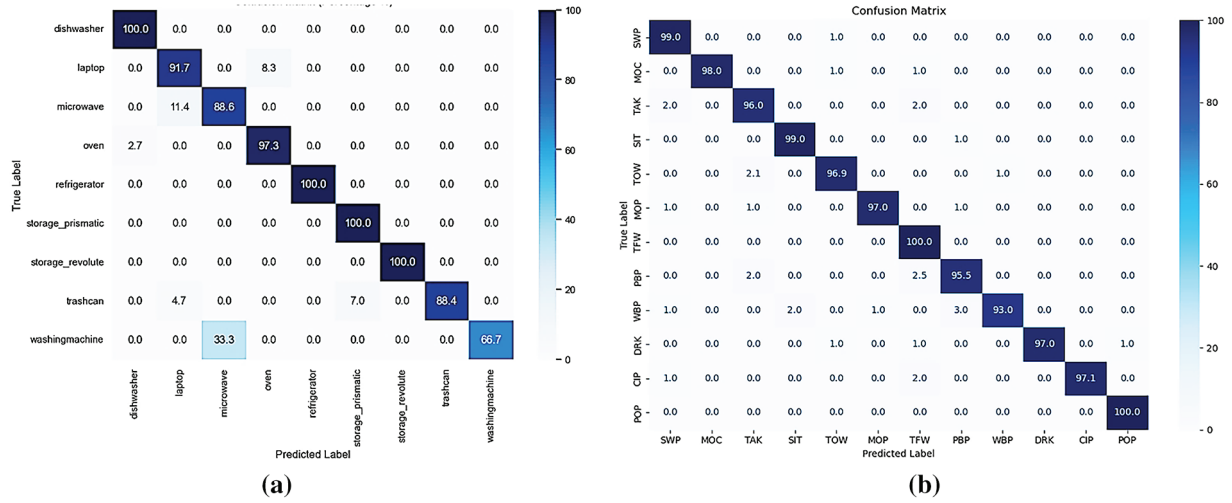
3D-D3D-HOI							Sysu 3D HOI Dataset						
Class	**Pr**	**Re**	**F1**	**Sp**	**FPR**	**LL**	**Class**	**Pr**	**Re**	**F1**	**Sp**	**FPR**	**LL**
Dishwasher	0.98	1.00	0.99	0.94	0.06	0.12	SWP	0.97	0.99	0.98	0.99	0.01	0.10
Laptop	0.79	0.92	0.85	0.91	0.09	0.29	MOC	1.00	0.98	0.99	1.00	0.00	0.00
Microwave	0.84	0.89	0.86	0.88	0.12	0.17	TAK	0.98	0.96	0.97	0.98	0.02	0.15
Oven	0.95	0.97	0.96	0.97	0.03	0.08	SIT	0.98	0.99	0.99	0.99	0.01	0.08
Refrigerator	1.00	1.00	1.00	1.00	0.00	0.00	TOW	0.98	0.98	0.98	0.98	0.02	0.12
S-Prismatic	0.84	1.00	0.91	0.99	0.01	0.10	MOP	1.00	0.99	0.99	1.00	0.00	0.00

(Continued)

Table 4 (continued)

3D-D3D-HOI							Sysu 3D HOI Dataset						
Class	Pr	Re	F1	Sp	FPR	LL	Class	Pr	Re	F1	Sp	FPR	LL
S- Revolute	1.00	1.00	1.00	1.00	0.00	0.00	TFW	0.96	1.00	0.98	0.95	0.05	0.18
Trashcan	1.00	0.88	0.94	1.00	0.00	0.00	PBP	0.96	1.00	0.98	1.00	0.00	0.00
Washing machine	1.00	0.67	0.80	0.33	0.67	0.45	WB	0.99	1.00	0.99	0.99	0.01	0.07
							DRK	1.00	0.94	0.97	1.00	0.00	0.00
							CIP	1.00	1.00	1.00	1.00	0.00	0.00
							POP	0.99	1.00	0.99	1.00	0.00	0.00

Note: SWP = sweeping, PBP = pocket backpacks, MOC = moving chair, TOW = taking out wallet, MOP = mop-ping, TFW = taking from wallet, WBP = wearing backpack, TAK = pouring, SIT = sitting in chair, CIP = calling phone, DRK = drinking, POP = playing on phone.



(a) (b)

Figure 7: Confusion matrix for model evaluation on the 3D-D3D-HOI (**a**) dataset and (**b**) Sysu 3d HOI dataset

The model's performance is shown by the AUC values, where higher classification performance is indicated by values near 1.00. The ROC curves in Fig. 8a,b demonstrate how well the model distinguishes between classes; the curves for the majority of classes clearly deviate from the diagonal, suggesting that the model successfully detects positive cases for each class.

Fig. 9a indicates that the model is successfully learning and minimizing errors over time. The training accuracy plots in Figs. 9b demonstrate the improvement in accuracy as training progresses.

Table 5 displays the performance of the QI-CNN model in detecting HOI scenarios from the 3D-D3D-HOI data and Sysu 3D HOI dataset. The table displays the loss and accuracy data for every epoch, ranging from epoch 1 to epoch 100. The loss decreases as more data is processed during training. In Table 6, we represent comparison analysis of our model with state-of-the-art methods.

We evaluated different models for HAR through testing on both D3D HOI. We conducted an ablation study comparing classical models and quantum-classical hybrid models, as well as our QI-CNN. Our analysis

involved studying both versions of the models, which included or excluded quantum gates, to measure changes in performance metrics.

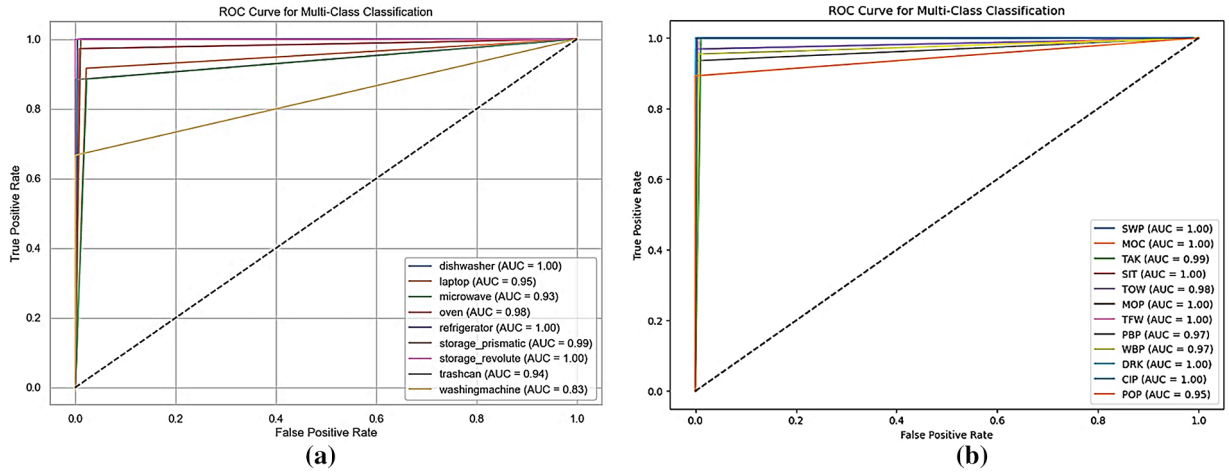


Figure 8: ROC curve for the (**a**) 3D-D3D-HOI dataset (**b**) Sysu 3d HOI dataset

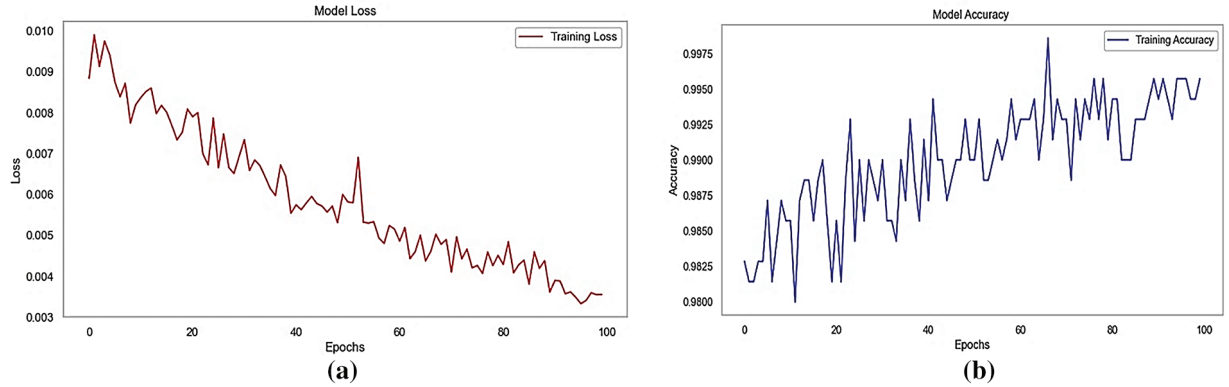


Figure 9: 3D-D3D-HOI dataset (**a**) training loss over epochs (**b**) training accuracy over epochs

Table 5: Results for Accuracy and Loss for the human interaction detection using the QI-CNN model's testing

3D-D3D-HOI dataset			Sysu 3d HOI dataset		
Epoch	**Loss**	**Accuracy**	**Epoch**	**Loss**	**Accuracy**
1	0.0088	0.9929	1	0.0573	0.9403
2	0.0099	0.9814	2	0.0565	0.9304
10	0.0082	0.9857	10	0.0482	0.9403
20	0.0081	0.9814	20	0.0405	0.9489
35	0.0064	0.9900	35	0.0315	0.9588
40	0.0055	0.9914	40	0.0320	0.9545
50	0.0060	0.9900	50	0.0265	0.9702
100	0.0035	0.9957	100	0.0157	0.9830

Table 6: Comparison of accuracy for HAR recognition with state of the art method

Methods	Year	SYSU Dataset	D3D HOI
CTRS-GCN [22]	2021	53.33	–
Multi-modality co-learning [23]	2024	80.83	–
K-ary key hashing [24]	2021	93.5	–
Fully Convolutional network [25]	2022	91.68	–
CAD Model [20]	2021	–	90.91
Implicit Field Fitting [26]	2022	–	74.3
Graph-Based Approach [27]	2022	–	89.6
Proposed Approach		97.38	93.02

Table 7 presents the test accuracy for both QI-CNN and CNN models across different object classifications. The values indicate the overall performance in terms of classification accuracy.

Table 7: Accuracy analysis of our proposed QI-CNN model 8 classes combination of D3D HOI dataset

Combination	Test accuracy (QI-CNN)	Test accuracy (CNN)
Dishwasher vs. Laptop	1.00	0.95
Oven vs. Microwave	0.97	0.93
Refrigerator vs. Washing Machine	1.00	0.83
Storage Revolute vs. Trashcan	1.00	0.94

Table 8 shows the performance analysis of both QI-CNN and CNN models across different combinations of object classes. It includes train, validation, and tests for both 1000 and 500 images per class in Table 9.

Table 8: Performance analysis of our proposed QI-CNN with the CNN model with 1000 images for 8 datasets classes

Accuracy	1000 images per class							
	Dishwasher vs. Laptop		Oven vs. Microwave		Refrigerator vs. Washing Machine		Storage Revolute vs. Trashcan	
	QI-CNN	CNN	QI-CNN	CNN	QI-CNN	CNN	QI-CNN	CNN
Train	94.50%	92.30%	96.00%	91.00%	100%	85.00%	100%	92.80%
Validation	92.00%	90.10%	94.50%	89.30%	99.80%	83.00%	99.90%	91.00%
Test	91.20%	89.30%	93.20%	88.50%	99.90%	80.30%	99.80%	90.20%

Table 9: Performance analysis of our proposed QI-CNN with CNN model with 500 images per class for 8 datasets

Accuracy	500 images per class							
	Dishwasher vs. Laptop		Oven vs. Microwave		Refrigerator vs. Washing machine		Storage Revolute vs. Trashcan	
	QI-CNN	CNN	QI-CNN	CNN	QI-CNN	CNN	QI-CNN	CNN
Train	93.00%	91.50%	94.30%	90.40%	98.00%	91.50%	99.50%	91.80%
Validation	91.80%	89.00%	93.10%	88.50%	96.70%	90.00%	98.10%	90.40%
Test	90.00%	88.40%	92.30%	87.90%	97.80%	89.00%	97.80%	89.90%

Table 10 presents the confidence intervals (CI) for the QuA-CNN and CNN models. The 95% confidence intervals are computed based on the performance of 1000 images per class.

Table 10: Confidence Intervals (95%) for QI-CNN and CNN Models for 8 datasets classes

Combination	QI-CNN CI (1000 Images)	CNN CI (1000 Images)
Dishwasher vs. Laptop	[85.45–90.30]	[84.33–88.56]
Oven vs. Microwave	[86.64–93.77]	[90.85–91.88]
Refrigerator vs. Washing machine	[88.01–92.45]	[81.91–88.32]
Storage revolute vs. Trashcan	[90.03–94.90]	[86.12–90.21]

5 Discussion

Our QI-CNN model has a quantum layer that employs quantum gates (RY, RX, RZ, CNOT, CZ) and amplitude encoding to improve feature mapping. These quantum operations allow more effectively rep-resent high-dimensional, complex data, and high-level features that cannot be extracted by classical CNNs, which is especially applicable to human interaction recognition. Our model enables the quantum and classical layers to interact dynamically. The quantum layer assists in mapping and optimization of features, and the classical CNN layers optimize the features. This refinement is adaptive and enhances optimization and enables the model to capture non-linear relationships and deal with uncertainty in the data. Our QI-CNN incorporates the quantum layer in a new manner compared to the current quantum-classical CNNs, which offer adaptive refinement during training. The earlier models rely on quantum layers to extract basic features, whereas our model improves the feature extraction and optimization, which is more efficient in complex and high-dimensional data such as human interaction recognition. The quantum circuit consists of 3 layers of parametrized quantum gates with 24 trainable parameters on 8 qubits. The circuit depth is about 20 gates. These parameters are optimized with the Adam optimizer (learning rate = 0.01, 500 iterations). Qiskit Aer is used to simulate the quantum part, and 1024 repetitions are used to simulate each measurement.

Table 11 represents the full model with all components (HoG, GIST, Skeleton Joints, and Quantum Layer) achieves the highest accuracy for D3D-HOI (93.02%) and Sysu 3D HOI (97.38%), with F1-Scores of 92.30% and 97.44%, respectively. This demonstrates the effectiveness of the combined classical-quantum approach in enhancing performance on both datasets. Without Quantum Layer: The exclusion of the quantum layer results in a significant drop in performance (−5.3% accuracy and −6.6% F1-Score) for both datasets, highlighting the value added by quantum optimization in feature transformation. Without HoG, GIST, or Skeleton Joints: While the classical features still provide a reasonable performance boost, their removal causes a decline in accuracy, with HoG contributing the most to overall performance, especially for D3D-HOI. HoG's removal shows the most significant impact on both accuracy and F1-Score across both datasets. When Fuzzy Optimization is removed, accuracy decreases by −1.77% for D3D-HOI and −2.04% for Sysu 3D HOI, underlining the importance of fuzzy logic in optimizing feature selection and handling ambiguity.

To ensure the reliability and fairness of our results, we conducted 10-fold cross-validation on the D3D-HOI datasets. Table 12 reports fold-wise performance across Accuracy, Precision, Recall, and F1-Score. The results demonstrate consistent performance across all folds, with low variance, indicating that our method is robust and not biased by a specific data split.

Table 11: Ablation study on feature contribution with quantum layer

Experiment	HoG	GIST	Skeleton joints	Fuzzy optimization	Quantum layer	D3D-HOI Acc (%)	Sysu 3D HOI Acc (%)	D3D-HOI F1-Score (%)	Sysu 3D HOI F1-Score (%)
Full Model (FM)	✓	✓	✓	✓	✓	93.02	97.38	92.30	97.44
without quantum layer	✓	✓	✓	✓	✗	87.52	92.50	85.67	91.12
without HoG	✗	✓	✓	✓	✓	89.23	93.60	87.65	94.02
without GIST	✓	✗	✓	✓	✓	88.94	92.40	87.14	93.30
without fuzzy optimization	✓	✓	✓	✗	✓	91.25	91.25	91.25	91.25
without skeleton joints	✓	✓	✗	✓	✓	90.12	94.10	88.41	94.55

Note: ✓ represent used technique in experiment and ✗ show excluded techniques.

Table 12: K-fold cross-validation results for 3D-D3D-HOI dataset

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	93.02	92.50	91.85	92.18
2	92.85	92.30	91.50	91.90
3	93.10	92.70	92.10	92.40
4	92.95	92.80	91.95	92.38
5	93.05	92.90	92.20	92.55
6	92.85	92.60	91.75	92.05
7	93.15	92.75	92.00	92.37
8	93.20	92.85	92.15	92.50
9	93.00	92.60	91.90	92.25
10	93.10	92.80	92.05	92.42
Mean ± Std	**93.02 ± 0.1**	**92.70 ± 0.15**	**92.00 ± 0.15**	**92.30 ± 0.15**

The tight standard deviation (σ = 0.1%) across all folds demonstrates exceptional stability, ensuring that the performance is consistent and not reliant on a specific split of the data. Cross-validation ensures that each model is evaluated on different subsets of the data, making the evaluation more robust and reliable than a single train-test split.

Table 13 compares the recognition accuracy, inference time, and GFLOPs (Giga floating-point operations) for computational efficiency of the QI-CNN model with various baseline models, including traditional CNN, HCNN, and quantum-enhanced models (QCCNN, HQC-CNN, and QC-CNN). The QI-CNN model demonstrates competitive accuracy while balancing computational cost with efficient feature extraction through its hybrid quantum-classical architecture.

Table 13: Comparison of classification performance and computational efficiency of the proposed QI-CNN model with other models

Model	Recognition accuracy (%)	Inference time (ms/frame)	GFLOPs
CNN	91.12	23.5	32
HCNN	92.78	25.0	35
QCCNN (No Gates)	93.80	27.2	40
HQC-CNN (No Gates)	94.50	28.5	42

(Continued)

Table 13 (continued)

Model	Recognition accuracy (%)	Inference time (ms/frame)	GFLOPs
QC-CNN	94.60	30.1	45
HQCCNN	93.80	29.5	43
QI-CNN (Ours)	93.02	32.0	48

6 Conclusion

This work proposes a novel Quantum-Integrated Convolutional Neural Network (QI-CNN) for human interaction recognition, integrating quantum computing with classical CNNs to enhance feature extraction and learning. The model achieved 93.02% accuracy on the D3D-HOI dataset and 97.38% accuracy on the SYSU 3D HOI dataset, demonstrating the potential of quantum-assisted models in complex human activity recognition tasks. Limitations of the proposed hybrid quantum-classical model include the constraints of current quantum software, such as qubit availability, quantum noise, and circuit depth. Additionally, the use of quantum simulators for complex datasets introduces computational overhead, despite the promising results. Nevertheless, the proposed approach offers significant potential for applications in surveillance, robotics, and assistive technologies. This work demonstrates that quantum computing can effectively enhance conventional CNNs, opening up avenues for future research into more efficient quantum encodings, better quantum hardware, and scalable methods to handle larger datasets. Future work will focus on scaling the quantum components, minimizing the circuit depth, and applying the model to larger and more diverse datasets. The development of quantum hardware with more qubits and reduced error rates will also improve the model's performance and may lead to the introduction of quantum-enhanced deep learning for practical, real-world applications.

Acknowledgement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding Statement: The APC was funded by the Open Access Initiative of the University of Bremen and the DFG via SuUB Bremen. This research was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project Number (PNURSP2025R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: Conceptualization, Tanvir Fatima Naik Bukht, Nouf Abdullah Almujally and Hameedur Rahman; methodology, Tanvir Fatima Naik Bukht and Hameedur Rahman; software, Tanvir Fatima Naik Bukht, Hameedur Rahman and Shuoa S. AItarbi; validation, Yanfeng Wu and Shuoa S. AItarbi; formal analysis, Yanfeng Wu, Tanvir Fatima Naik Bukht and Hameedur Rahman; investigation, Nouf Abdullah Almujally; resources, Shuoa S. AItarbi; data curation, Tanvir Fatima Naik Bukht; writing—original draft preparation, Tanvir Fatima Naik Bukht, Ahmad Jalal and Hui Liu; writing—review and editing, Yanfeng Wu, Nouf Abdullah Almujally and Hameedur Rahman; visualization; supervision, Hameedur Rahman and Ahmad Jalal; project administration, Hameedur Rahman, Ahmad Jalal and Tanvir Fatima Naik Bukht; project; funding acquisition, Hameedur Rahman, Ahmad Jalal and Hui Liu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets analyzed during the current study are available in the [D3D-HOI] repository, [https://github.com/facebookresearch/d3d-hoi] (accessed on 02 November 2024) and [SYSU_3D_HOI] [https://opendatalab.com/OpenDataLab/SYSU_3D_HOI_Set] (accessed on 06 March 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Dharmendrasinh R, Thakkar A, Parmar D, Patel K. Human object interaction: a survey on models and their key challenges and potential applications in future fields. In: Proceedings of the International Conference on Artificial Intelligence and Smart Energy; 2024 Mar 22–23; Coimbatore, India. p. 93–106.
2. Manakitsa N, Maraslidis GS, Moysis L, Fragulis GF. A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. Technologies. 2024;12(2):15. doi:10.3390/technologies12020015.
3. Liu S, Li YL, Fang Z, Liu X, You Y, Lu C. Primitive-based 3D human-object interaction modelling and programming. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2024 Feb 26–27; Vancouver, BC, Canada. p. 3711–9.
4. Li Q, Xie X, Zhang C, Zhang J, Shi G. Detecting human-object interactions in videos by modeling the trajectory of objects and human skeleton. Neurocomputing. 2022;509:234–43. doi:10.1016/j.neucom.2022.08.008.
5. Meedinti GN, Srirekha KS, Delhibabu R. A quantum convolutional neural network approach for object detection and classification. arXiv:2307.08204. 2023.
6. Liu J, Lim KH, Wood KL, Huang W, Guo C, Huang HL. Hybrid quantum-classical convolutional neural networks. Sci China Phys Mech Astron. 2021;64(9):290311. doi:10.1007/s11433-021-1734-3.
7. Ranga D, Rana A, Prajapat S, Kumar P, Kumar K, Vasilakos AV. Quantum machine learning: exploring the role of data encoding techniques, challenges, and future directions. Mathematics. 2024;12(21):3318. doi:10.3390/math12213318.
8. Ling YQ, Zhang JH, Zhang LH, Li YR, Huang HL. Image classification using hybrid classical-quantum neutral networks. Int J Theor Phys. 2024;63(5):125. doi:10.1007/s10773-024-05669-w.
9. Ren Z, Zhang Q, Gao X, Hao P, Cheng J. Multi-modality learning for human action recognition. Multimed Tools Appl. 2020;80(11):16185–203. doi:10.1007/s11042-019-08576-z.
10. Hou Z, Peng X, Qiao Y, Tao D. Visual compositional learning for human-object interaction detection. In: Proceedings of the Computer Vision—ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK. p. 584–600. doi:10.1007/978-3-030-58555-6_35.
11. Zhang FZ, Campbell D, Gould S. Spatially conditioned graphs for detecting human-object interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 13319–27.
12. Kim S, Jung D, Cho M. Relational context learning for human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 2925–34.
13. Park J, Park JW, Lee JS. ViPLO: vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 17152–62. doi:10.1109/CVPR52729.2023.01645.
14. Quan H, Lai H, Gao G, Ma J, Li J, Chen D. Pairwise CNN-transformer features for human-object interaction detection. Entropy. 2024;26(3):205. doi:10.3390/e26030205.
15. Li R, Zeng X, Yang S, Li Q, Yan A, Li D. ABYOLOv4: improved YOLOv4 human object detection based on enhanced multi-scale feature fusion. EURASIP J Adv Signal Process. 2024;2024(1):6. doi:10.1186/s13634-023-01105-z.
16. Gavali AS, Kakarwal SN. Enhancing early action prediction in videos through temporal composition of sub-actions. Multimed Tools Appl. 2025;84(8):4253–81. doi:10.1007/s11042-024-18870-0.
17. Zhang S, Ma Z, Zhang G, Lei T, Zhang R, Cui Y. Semantic image segmentation with deep convolutional neural networks and quick shift. Symmetry. 2020;12(3):427. doi:10.3390/sym12030427.
18. Tang HH, Ahmad NS. Fuzzy logic approach for controlling uncertain and nonlinear systems: a comprehensive review of applications and advances. Syst Sci Control Eng. 2024;12(1):2394429. doi:10.1080/21642583.2024.2394429.
19. Zheng Y, Xu Z, Wang X. The fusion of deep learning and fuzzy systems: a state-of-the-art survey. IEEE Trans Fuzzy Syst. 2021;30(8):2783–99. doi:10.1109/tfuzz.2021.3062899.

20. Xu X, Joo H, Mori G, Savva M. D3D-hoi: dynamic 3D human-object interactions from videos. arXiv:2108.08420. 2021.
21. Hu JF, Zheng WS, Lai J, Zhang J. Jointly learning heterogeneous features for RGB-D activity recognition. IEEE Trans Pattern Anal Mach Intell. 2017;39(11):2186–200. doi:10.1109/cvpr.2015.7299172.
22. Chen Y, Zhang Z, Yuan C, Li B, Deng Y, Hu W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, BC, Canada. p. 13359–68.
23. Liu J, Chen C, Liu M. Multi-modality co-learning for efficient skeleton-based action recognition. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct 28–Nov 1; Melbourne, Australia. p. 4909–18.
24. Khalid N, Ghadi YY, Gochoo M, Jalal A, Kim K. Semantic recognition of human-object interactions via gaussian-based elliptical modeling and pixel-level labeling. IEEE Access. 2021;9:111249–66. doi:10.1109/ACCESS.2021.3101716.
25. Ghadi YY, Waheed M, Al Shloul T, Alsuhibany A, Jalal S, Park A, et al. Automated parts-based model for recognizing human-object interactions from aerial imagery with fully convolutional network. Remote Sens. 2022;14(6):1492. doi:10.3390/rs14061492.
26. Haresh S, Sun X, Jiang H, Chang AX, Savva M. Articulated 3D human-object interactions from RGB videos: an empirical analysis of approaches and challenges. In: Proceedings of the 2022 International Conference on 3D Vision (3DV); 2022 Sep 12–15; Prague, Czechia, New York, NY, USA: IEEE; 2022. doi:10.1109/3DV57658.2022.00043.
27. Ghadi YY, Waheed M, Gochoo M, Alsuhibany SA, Chelloug SA, Jalal A, et al. A graph-based approach to recognizing complex human-object interactions in sequential data. Appl Sci. 2022;12(10):5196. doi:10.3390/app12105196.