Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.072243

ARTICLE

Hybrid Quantum Gate Enabled CNN Framework with Optimized Features for Human-Object Detection and Recognition

Nouf Abdullah Almujally[1], Tanvir Fatima Naik Bukht[2], Shuaa S. Alharbi[3], Asaad Algarni[4], Ahmad Jalal[2,5] and Jeongmin Park[6,*]

[1]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11564, Saudi Arabia
[2]Department of Computer Science, Air University, E-9, Islamabad, 44000, Pakistan
[3]Department of Information Technology, College of Computer, Qassim University, Buraydah, 52571, Saudi Arabia
[4]Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha, 91911, Saudi Arabia
[5]Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul, 02841, Republic of Korea
[6]Department of Computer Engineering, Tech University of Korea, 237 Sangidaehak-ro, Siheung-si, 15073, Republic of Korea

*Corresponding Author: Jeongmin Park. Email: jmpark@tukorea.ac.kr

Received: 22 August 2025; Accepted: 28 November 2025; Published: 10 February 2026

ABSTRACT: Recognising human-object interactions (HOI) is a challenging task for traditional machine learning models, including convolutional neural networks (CNNs). Existing models show limited transferability across complex datasets such as D3D-HOI and SYSU 3D HOI. The conventional architecture of CNNs restricts their ability to handle HOI scenarios with high complexity. HOI recognition requires improved feature extraction methods to overcome the current limitations in accuracy and scalability. This work proposes a Novel quantum gate-enabled hybrid CNN (QEH-CNN) for effective HOI recognition. The model enhances CNN performance by integrating quantum computing components. The framework begins with bilateral image filtering, followed by multi-object tracking (MOT) and Felzenszwalb superpixel segmentation. A watershed algorithm refines object boundaries by cleaning merged superpixels. Feature extraction combines a histogram of oriented gradients (HOG), Global Image Statistics for Texture (GIST) descriptors, and a novel 23-joint keypoint extraction method using relative joint angles and joint proximity measures. A fuzzy optimization process refines the extracted features before feeding them into the QEH-CNN model. The proposed model achieves 95.06% accuracy on the 3D-D3D-HOI dataset and 97.29% on the SYSU 3D HOI dataset. The integration of quantum computing enhances feature optimization, leading to improved accuracy and overall model efficiency.

KEYWORDS: Pattern recognition; image segmentation; computer vision; object detection

1 Introduction

Human–object interaction (HOI) detection aims to localise people and objects in a scene and to classify the interaction between them. Success depends on accurate localisation and classification, and current approaches fall into categories such as two-stage detectors and transformer-based networks. A recent survey reports that two-stage and one-stage detectors dominate the field and that HOI recognition is challenged by long-tail interaction distributions and occlusion, which motivate the exploration of novel architectures [1]. The egocentric dataset Ego-HOIBench further highlights that first-person HOI recognition suffers from hand–object occlusions; its benchmark comprises more than 27k images with hand-verb-object annotations

and demonstrates that third-person detectors struggle on egocentric data [2]. Real-time HOI systems, such as the LightSTATE framework, use efficient frame sampling and a vision–language model to achieve high precision, recall, and F1 (~0.98) while processing frames within approximately 0.8–0.9 s [3]. These findings underscore the importance of designing efficient, occlusion-aware HOI models and motivate our hybrid quantum–classical approach. Recent HOI detectors employ diverse architectures. Two-stage pipelines detect humans and objects separately before classifying interactions [4]. Despite these advances, the 2025 survey notes that current methods still struggle with occlusions and long-tail interaction distributions [5]. Real-time frameworks like Lightweight demonstrate that efficient video sampling and vision–language models can achieve high precision and F1 scores while maintaining low latency [6].

This paper proposes a hybrid Quantum Gate-Enabled Convolutional Neural Network (QEH-CNN) that integrates quantum amplitude encoding and parameterised single- and two-qubit gates into a CNN. Before entering the network, input frames are denoised using a bilateral filter and segmented with multi-object tracking, Felzenszwalb superpixels, and a watershed algorithm. Features are extracted by combining a histogram of oriented gradients (HOG), the GIST descriptor, and a 23-joint skeleton representation that captures relative joint angles and proximities. These visual and skeletal features are optimised using fuzzy logic and then encoded into quantum states. The resulting QEH-CNN performs convolutional feature learning in a quantum–classical loop and ends with a softmax classifier. Evaluated on the D3D-HOI and SYSU 3D HOI datasets, the model achieves 95.06% and 97.29% recognition accuracy, respectively, demonstrating both improved generalisation and computational efficiency. The key findings of this work are the following:

- Implement a novel QEH-CNN architecture to identify the HOI in multidimensional data.
- Quantum integrating into a traditional CNN enhances HOI detection.
- A feature extraction approach that integrated both HOG and GIST, as well as new skeleton-based and Visual-based features, to extract interesting information in the HOI recognition process.
- Incorporation of quantum computing units, such as amplitude encoding and quantum gates (RY, RX, RZ, CNOT, CZ) into the CNN architecture can also be used to enhance feature mapping.
- Experimental verification demonstrates that QEH-CNN exhibits a significant degree of scalability when applied to large HOI datasets that contain complex and diverse interaction behaviors.

The rest of the paper is structured in the following way. Section 2 is the literature review on HOI recognition, deep learning models, and hybrid quantum networks. Section 3 outlines the system methodology, which consists of preprocessing, segmentation, feature extraction, and the QEH-CNN architecture. Section 4 displays the experimental design, data and findings. Section 5 is the conclusion of the paper and the discussion of the future directions.

2 Literature Review

The research field has now focused on adding temporal evaluation methods alongside framework understanding to develop HOI models. The combination of CNNs and recurrent neural networks (RNNs) or transformers in [7] enables the effective capture of temporal dependencies, which is helpful for video-based HOI recognition tasks. These approaches face two main limitations in scaling across datasets and handling complex real-world interactions. The field has moved toward general deep learning methods for HOI detection. Han et al.'s survey shows how deep learning techniques, particularly CNNs, improve generalization abilities. The authors note that pretrained object detection networks are used for generalization but show limited scalability due to low understanding of new object interactions in complex scenarios [5]. The researchers Yang et al. developed F-HOI as a framework that delivers precise semantic alignment systems for recognizing human-object interactions in three dimensions. The proposed model addresses previous

generalization problems by focusing on the challenges of interaction diversity and the requirement for models to detect previously unobserved objects. According to the study, focusing on enhancing the matching of people and images is crucial for achieving better scalability and accuracy [8].

Hafeez et al. [9] propose a hybrid quantum-classical neural network (H-QNN) for binary image classification. Uses a two-qubit quantum circuit integrated with a classical convolutional architecture. Addresses overfitting with small datasets and leverages the strengths of both quantum and classical approaches for scalable image classification. Limited by NISQ devices, which have few qubits and high quantum noise. Wang et al. [10] developed a shallow hybrid quantum-classical convolutional neural network for image classification. Utilizes kernel encoding, variational quantum circuits (VQC), and mini-batch gradient descent for training. Training with a shallow network and mini-batch gradient descent. Initially, traditional machine learning for felting HOI operated through manual feature creation while labeling data for object detection. Support vector machines (SVMs), decision trees and random forests operated on image-based visual features to classify human-object interaction categories. The models struggled when handling diverse interaction variations. Machine learning models proved challenging with dynamic interactions, as scenarios with object occlusions, different poses, and numerous objects posed difficulties.

Current models struggle to handle large-scale data efficiently. Quantum computing combined with CNNs can address scalability and optimization issues by leveraging fast information processing. Research shows that quantum optimization and feature mapping can improve performance [11]. Quantum gates and amplitude encoding in CNNs enable rapid, accurate analysis of complex data. Although QML is emerging, the methods in [12] demonstrate its potential to surpass classical optimization limits for HOI recognition. Quantum parallelism reduces training time and boosts deep learning performance.

Visual Compositional Learning (VCL) [13] breaks down HOI representations into object-specific and verb-specific features, and solves the long-tail distribution issue by allowing such features to be shared across samples of HOIs. This decomposition increases generalization, especially when there are imbalanced categories of interaction in the dataset. VCL however, has difficulties with dealing with interactions between ambiguous or overlapping categories of objects, which may result in misclassification. Ren et al. [14] suggested a multi-modality fusion algorithm that uses RGB and depth images to recognize human actions. Their method recorded the state-of-the-art results on several datasets using depth information to improve the interpretation of human actions. Although successful, the approach is limited in the way it can deal with occlusion and motion in RGB-D sequences, which can adversely affect its performance in challenging situations. Zhang et al. [15]. Presented the Spatially Conditioned Graph (SCG) to detect HOI, using graph-based approaches to describe interactions. Their model performed well on the HICO-DET dataset because it was able to capture spatial relations between humans and objects. The model, however, performs much worse when pose estimation is of low quality, which means that pose information is essential in HOI detection. Vision Transformer (ViT) backbone with Masking with Overlapped Area (MOA) module and pose-conditioned self-loop graph ViPLO was proposed by Park et al. as a two-stage HOI detector. This strategy gave a +2.07 mAP gain on the HICO-DET dataset by successfully capturing human pose data. Nevertheless, the approach is costly in terms of efficiency and scalability as it needs a lot of computational resources to train the ViT model [16].

Kim et al. [17] proposed a three-branch network, Multiplex Relation Embedding Network (MUREN), which is an HOI detection network that uses multiplex relation context. MUREN demonstrated state-of-the-art results on HICO-DET and V-COCO data sets by improving the transfer of contextual information across branches. Though successful, the approach has a restricted context transfer between branches within past approaches, which can influence the depth of relational reasoning. Li et al. [18] suggested a better YOLOv4 architecture (ABYOLOv4) that has multi-scale feature fusion, trained to detect human-objects. ABYOLOv4

had a 0.5% higher Average Precision (AP) and 45.3 MB smaller model size, providing a more efficient HOI detector. Nonetheless, it is poor in extreme occlusion, small target sizes in complicated backgrounds, and this underscores the importance of powerful detection mechanisms. Quan et al. [19] proposed a two-step approach to HOI detection based on CNN and Transformer features to achieve higher accuracy. Their method increased the interaction detection using the strengths of CNN and Transformer, and they demonstrated a competitive result on the HICO-DET and V-COCO datasets. However, there are effects of occlusion and misclassification of the objects on performance, which implies that more can be done. Gavali and Kakarwal [20] suggested a way to break down actions into local actions to enhance the accuracy of early prediction. By viewing only 40% of the video frames, their method was found to be quite accurate (82.5% on SYSU 3D HOI and 90% on MSR Daily Activity). It is, however, limited in its ability to deal with similar actions visually and needs temporal decomposition, which may complicate the detection process.

3 Materials and Methods

3.1 Proposed System

The proposed HOI recognition method combines CNNs with quantum computing technology for improved accuracy and efficiency. The model integrates classical algorithms and quantum computing approaches to detect objects and segmental features while optimizing performance. The dataset undergoes preprocessing with a bilateral filter for data refinement. The detection process uses MOT and superpixels with watershed detection to extract joint features from the skeleton model and visual descriptors, including HOG and GIST. The extracted features undergo optimization through fuzzy logic. A QEH-CNN model operates through amplitude encoding, data encoding, and single and two-qubit gates with quantum-rich transformations and feature-enlargement layers. Convolutional operations for feature learning follow the quantum layer functions, culminating in a softmax for classification. The combination of quantum computing with traditional deep learning enhances the precision and speed of object detection, as shown in Fig. 1.

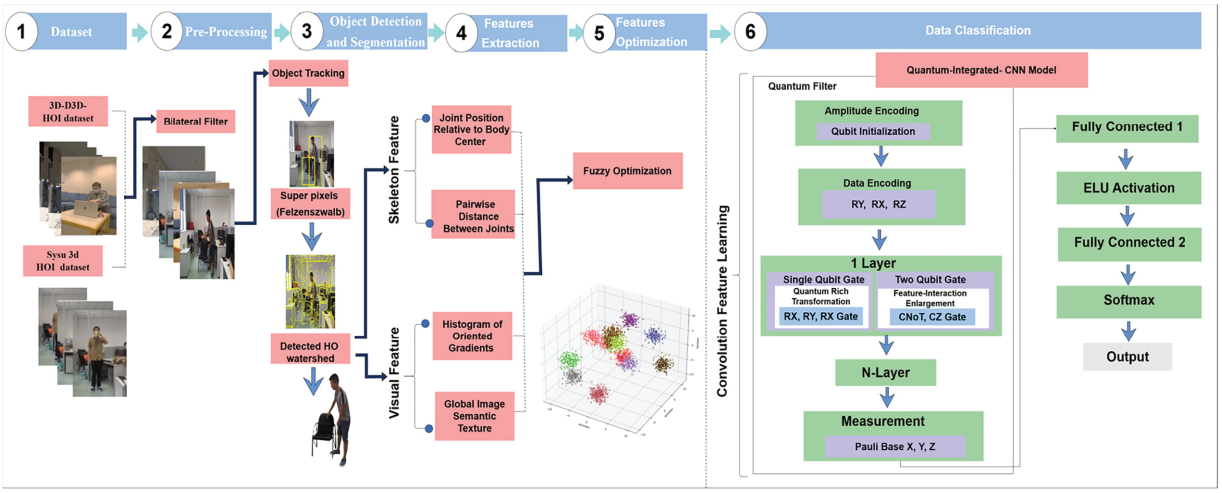


Figure 1: The proposed methodology demonstrates the combination of quantum computing and CNNs

3.2 Preprocessing

The original images have to be pre-processed to improve them and then they are used in further steps. BF is employed to remove noise and preserve the edges of images. BF is a nonlinear filter of images that

smoothes areas, depending on the proximity of pixels to each other and the difference in brightness between them. The bilateral filter consists of a weighting function and an averaging calculation which makes use of the weights. The filtered value $(\tilde{I}(p))$ is an outcome of a weighted summation process that is normalized.

$$\tilde{\boldsymbol{I}}(\boldsymbol{p}) = \frac{\mathbf{1}}{\sum_{\boldsymbol{q}\in\Omega} \boldsymbol{w}(\boldsymbol{p},\boldsymbol{q})} \sum_{\boldsymbol{q}\in\Omega} \boldsymbol{w}(\boldsymbol{p},\boldsymbol{q})\, \boldsymbol{I}(\boldsymbol{q}) \tag{1}$$

Eq. (1) computes pixel output $\tilde{I}(p)$ using neighboring intensities $I(q)$ within Ω, with normalization ensuring weights sum to one. The bilateral filter performs edge-preserving smoothing by combining spatial proximity and intensity similarity, reducing noise while retaining edges.

3.3 Object Detection and Segmentation

The process of segment refinement and HOI tracking is organized in several stages. MOT, superpixel segmentation with felzenszwalb, superpixel merging, and HO Watershed Segmentation. Each video is sampled at 30 frames per second. We extract sliding clips of 16 consecutive frames (≈0.53 s) with a stride of eight frames, resulting in 50% overlap between adjacent clips. Each clip inherits the action label of its source video. To aggregate frame-level predictions into a clip-level score, we compute the average of the softmax probabilities across all 16 frames. Multi-object tracking (MOT) is used solely to maintain consistent bounding boxes across frames; the MOT identifiers do not influence class labels.

3.3.1 Multi-Object Tracking (MOT)

The object detection pipeline begins with MOT, which detects and tracks multiple objects, including humans, across consecutive video frames. Unlike single object tracking (SOT), which tracks only one object, MOT handles complex scenarios like object identification, appearance changes, and occlusions. MOT uses bounding boxes to define regions of interest (RoIs), enabling precise localization and tracking as objects move, and ensures consistent tracking across frames, as shown in Fig. 2.

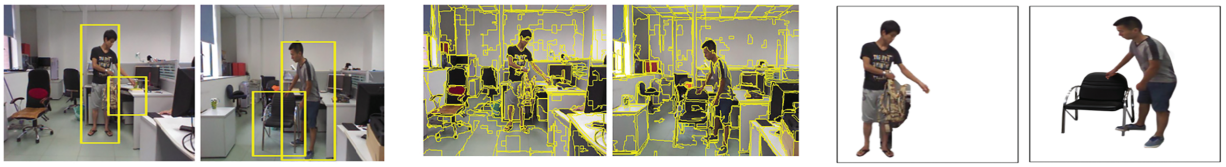


Figure 2: The figure shows the object detection and segmentation results, the results of MOT. (**a**) Carrying a backpack, (**b**) moving chair, the results of Felzenszwalb (**c**) and (**d**), the watershed of the detected human object presented in (**e**) and (**f**) beforehand

3.3.2 Superpixels (Felzenszwalb)

Grouping pixels into superpixels, the felzenszwalb algorithm exhibits common characteristics like color, texture, or intensity levels. When superpixels represent an image, they become less complex, making image segmentation and recognition easier. The analysis of HOI in images is made possible by Felzenszwalb's technique for superpixel segmentation and watershed detection [21]. Felzenszwalb's algorithm reduces the normalized cut algorithm to classify pixels into regions based on similar textures and colors. The method processes images rapidly while identifying edges in areas with varying textures and structures. After segmentation, watershed detection determines how people interact with objects in the image.

3.3.3 Detected HO Watershed

The final step in the segmentation process applies the HO watershed algorithm. Watershed segmentation treats the image as a topographic surface, where the pixels represent elevations, and the algorithm identifies boundaries between different regions. This technique is particularly effective in segmenting overlapping or touching objects, which is a common occurrence in HOI scenes. The HO 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.

3.4 Feature Extraction

It converts the pixel values of an image into useful representations that can be used in the classification process. The study aims to extract two basic features, such as visual-specific and skeleton-based features.

3.4.1 Visual Feature Extraction

The visual features are focused on the texture, shape and structure that assists in the analysis of the relationship that human beings have with objects in various situations. The HoG algorithm is capable of obtaining data on local edges and object outlines through the analysis of the gradient of x and y.

Global Image Semantic Texture (GIST)

GIST is a feature description that is used globally to explain the overall spatial structure and texture of a visual. It assists in determining the international backdrop of a visual. GIST is calculated with the help of a set of spatial filters of different scales and orientations that reflect the large-scale spatial and textual features of an image. GIST uses oriented filters of various sizes. The filters are given as Eq. (2).

$$\psi(x, y) = \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \cdot \cos\left(2\pi f\left(x\cos\theta + y\sin\theta\right)\right) \tag{2}$$

where $((x, y))$ are the pixel coordinates, (σ) is the standard deviation, (f) is the frequency, and (θ) is the orientation of the filter. In Eq. (2), the filter bank is used to take the responses of the image in varying spatial frequencies and orientations by convolving it with each filter in the bank. HoG pays attention to local gradient patterns, and GIST is interested in global spatial and texture patterns [22].

Histogram of Oriented Gradients (HoG)

HoG is one of the most essential feature descriptor techniques used in computer vision applications. The edge and gradient structures in local image regions are well detected by HoG to be used in object detection. One of the main ideas in HoG is that the local shapes of objects can be characterized by examining the distributions of gradient orientations in the regions around them.

$$G(x, y) = \sqrt{\left(\frac{\partial I(x, y)}{\partial x}\right)^2 + \left(\frac{\partial I(x, y)}{\partial y}\right)^2} \tag{3}$$

where Eq. (3) $(I(x, y))$ is the intensity of the image at pixel $((x, y))$. The gradient magnitude (G) and orientation (θ) at each pixel are then computed. The image is divided into small blocks or cells of size (8×8) pixels. For each cell, we compute a histogram of gradient orientations, which is quantized into bins, ranging from 0° to 180°. The resulting GIST and HoG descriptor captures the gradient distribution and is robust to changes in lighting and small spatial transformations, as shown in Fig. 3.

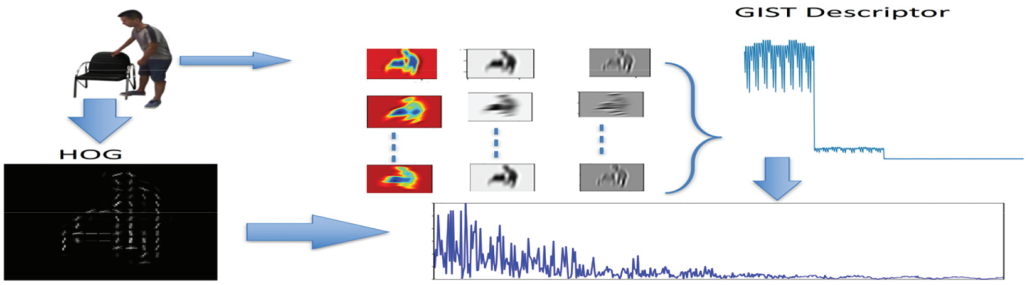
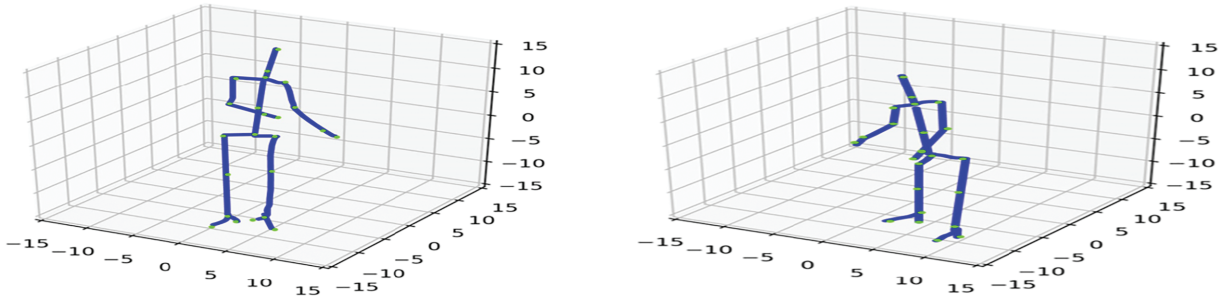


Figure 3: The figure illustrates the feature extraction results of HOG and GIST

3.4.2 Skeleton Feature Extraction

The skeletal geometry features are useful in cases where the information is obtained using the skeletal structure to detect the movement and behavior of humans. Motion-capture systems can be developed based on skeleton geometry features, thereby creating a significant amount of information about the human body structure and motion. They could be used in numerous other activities, including action recognition, gait analysis, and human-computer interaction. Human and object interaction with the new skeleton model, in which the 23 points of importance (green) and the skeleton interconnections (blue) are superimposed on Fig. 4. and 3D representation of the same skeleton model, in which the key joint points and overall structure are identified. The model is used in accurate human pose tracking in interaction recognition.



(**a**) (**b**)

Figure 4: Illustrate 3D visualization of the skeleton model with the key joints. (**a**) Wearing a backpack; (**b**) Moving chair

Joint Position Relative to Body Center

To standardize joint positions across different subjects or poses, we use statistical matching instead of traditional geometric transformations. We align these joint positions by minimizing the statistical discrepancy, using a linear transformation based on covariance. The objective function is given by:

$$\text{minimize}(A, b) \sum_{i=1}^{N} \left(|Ap_i + b - q_i|_2^2 \right) \tag{4}$$

where in Eq. (4) (A) is a transformation matrix and (b) is a translation vector. The resulting normalized joint positions are: $\widehat{p_i} = Ap_i + b$. This approach removes the effects of translation, rotation, and scaling by statistically aligning the poses, making them comparable across different subjects.

Pairwise Distance between Joints

Instead of using Euclidean distances, we analyze the human skeleton using a graph model $G = (V, E)$, where in Eq. (5) V are the joints and E represent bones connecting these joints. The geodesic distance between two joints i and j is the shortest path in the graph, considering anatomical connectivity rather than direct spatial distances. Where path (i, j) is the shortest path between joints (i) and (j), and (M) is the number of edges traveled.

$$d_G(i, j) = \min_{\text{path}(i,j)} \sum_{k=1}^{M} |p_{u_k} - p_{v_k}| \tag{5}$$

3.5 Feature Optimization

The process of fuzzy optimization begins by assigning a degree of relevance to each feature as a fuzzy membership value. Here, we have a membership function that describes how strongly a particular feature should weigh in when deciding, with one being closer to its maximum possible relevance. A common approach is to use a smooth and continuous mapping performed by a sigmoid membership function between the feature's value and its score in significance.

$$\mu(F_i) = \frac{1}{1 + e^{-\alpha(F_i - \mu_0)}} \tag{6}$$

Eq. (6) can be represented as the membership function for feature F_i. The fuzzy value $\mu(F_i)$ with parameter α controls the transition between relevant and irrelevant features, while μ_0 sets the relevance threshold. Using a sigmoid function, fuzzy optimization refines HOI feature selection, handling uncertain posture/activity data and improving decision-making. Combined with fuzzy inference and selection techniques, it boosts recognition accuracy, reduces redundant data, and maintains efficiency. Fig. 5 illustrates feature discrimination on Sysu 3D HOI. Fuzzy logic is better than PCA and ICA since it allows object features and data instances to belong to more than one class to a certain extent, thereby providing a more realistic description of natural human actions. Human-object interactions are a special case of recognition that needs special consideration of human poses and object occlusions and partial interactions, as these aspects often blur the boundaries of classes.

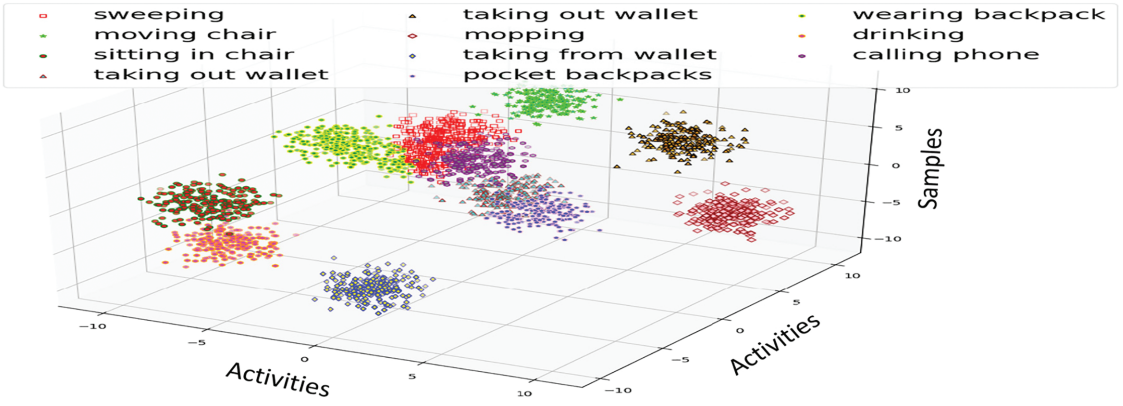


Figure 5: Illustration of fuzzy optimization on extracted features of Sysu 3D HOI dataset

3.6 Quantum Gate-Enabled Hybrid Convolutional Neural Network Model

The QEH-CNN is a new model of human activity recognition, which is advantageous in merging the virtues of quantum computing and classical deep learning methods. This hybrid design builds on the capability of quantum computing to compute high-dimensional data and the capability of classical CNN to learn powerful features and classify them. The model is performed in two major steps: quantum step to transform features and the classical step to learn features and classify activities.

Fig. 6 represents the qubit quantum circuit architecture of QEH-CNN, in which one layer of single-qubit rotations (R Y gates) followed by a ring of CNOT gates between the qubits. The circuit is made up of 3 layers of parameterized qubit rotations, and the rotation angle (0) is a trainable parameter. A ring topology of CNOT gates is used to entangle the qubits in an entanglement pattern to enable global interactions of features. The circuit is then measured by measuring the expectation values of Pauli-Z on each qubit after the transformations and the results are a transformed feature vector. Three layers of single-qubit RY rotations like trainable angles and ring CNOTs, are used; there are eight qubits and 24 parameters; the circuit depth is ~20 gates. Include the simulator description (Qiskit Aer state-vector with 1024 shots, double precision), initialisation and optimisation settings. The three ablation variants and state that all are trained under identical hyperparameters except for the presence/absence of quantum gates.

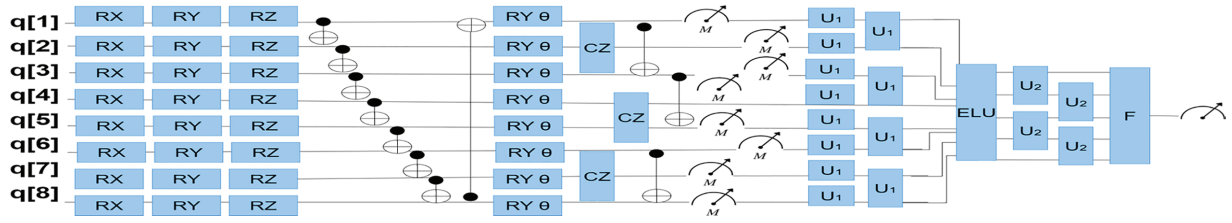


Figure 6: Hybrid convolutional neural network model with quantum gate

3.6.1 Quantum Filter and Feature Transformation

In the QEH-CNN model, the first step is to take classical features and translate them into a format that can be understood by quantum systems. We use a method called amplitude encoding, which normalizes the feature vector and maps it to a quantum system. The process looks like this Eq. (7).

$$|\varphi\rangle = \frac{1}{|X|_2}\sum_{i=1}^{n} X_i\,|q_i\rangle\,,\ \text{where}\ |X|_2 = \sqrt{\sum_{i=1}^{n}|X_i|^2} \tag{7}$$

Encoding the features this way, we turn them into a quantum state, ($|\varphi\rangle$), which exists in a high-dimensional quantum space. This gives the system the ability to process and understand the relationships between features much more powerfully than traditional methods. Because amplitude encoding represents a classical vector φ as a normalised quantum state $|\varphi\rangle$ of n qubits only when the length of x is a power of two, we compress our fused HOG, GIST, skeleton feature vector to 256 elements and pad it with zeros if necessary. This 256-dimensional vector is then normalised and loaded into an eight-qubit register using Qiskit's amplitude-embedding routine, which performs the required padding and normalisation automatically.

3.6.2 Quantum Gate Operation

Once the data is encoded into quantum states, we need to apply quantum gates to refine these features. These gates like (R_x), (R_y) and (R_z) work by altering the states of individual qubits. The operations happen

across several layers, and each layer involves rotating qubits and entangling them to create more complex relationships between the features. The overall process is represented by Eq. (8):

$$|\varphi_{\text{out}}\rangle = \prod_{l=1}^{L} \left(\bigotimes_{k=1}^{m} R_k(\theta_l) \right) \cdot E_l |\varphi\rangle \tag{8}$$

here, $(R_k(\theta_l))$ refers to the rotation on the (k^{th}) qubit in the (l^{th}) layer, and (E_l) describes how qubits are entangled in that layer. These layers of quantum gates help the system explore deeper feature interactions and uncover hidden patterns in the data. Two-qubit gates like CNOT also introduce entanglement, allowing the system to process more complex feature relationships. The encoded state is processed by a parametrised circuit of three layers. Each layer applies rotations around the y-axis to each of the eight qubits and then a ring of controlled-NOT gates to entangle them. This ansatz has twenty-four trainable parameters and a depth of roughly twenty gates. The circuit parameters and the weights of the classical CNN are trained together using the Adam optimiser with a learning rate of 0.01, and gradients are computed using the parameter-shift rule instead of back-propagation. We use the state-vector simulator in Qiskit Aer for all quantum operations. The input feature vector is first reduced to 256 dimensions, 2n and normalised. It is then amplitude-encoded into an eight-qubit register. The variational circuit consists of three layers. Each layer applies single-qubit rotations $R_y(\theta_{k,l})$ to every qubit followed by a ring of CNOT gates that entangle qubit k with qubit $k + 1$, wrapping the last qubit back to the first. After the final layer, expectation values of Pauli-Z observables are measured to produce the quantum-enhanced features. Algorithm 1 is as follows:

Algorithm 1: Description of the quantum component

```
Input: classical feature vector X
X ← PCA_compress(X, 256)
x_norm ← normalise_and_pad( X)
state ← amplitude_encode(x_norm) # 8 qubits
for l = 1 to 3:
    for k = 1 to 8:
        apply R_y(θ_{k,l}) on qubit k
    for k = 1 to 7:
        apply CNOT(k → k + 1)
    apply CNOT(8 → 1) # ring entanglement
return measure_Z(state)
```

3.6.3 Classical CNN for Feature Learning and Classification

Once the quantum system has transformed the features, the classical CNN takes over to do the feature learning and classification. The quantum-enhanced features (X_{quantum}) are passed through the first fully connected layer of the CNN. $Z_1 = W_1 \cdot X_{\text{quantum}} + b_1$ here, (W_1) is the weight matrix, and (b_1) is the bias. The output (Z_1) goes through an activation function like ReLU or ELU to introduce some non-linearity into the system, $Z_1' = \text{ELU}(Z_1)$. Next, the transformed features (Z_1') move on to the second fully connected layer. $Z_2 = W_2 \cdot Z_1' + b_2$. Finally, the output layer applies the softmax activation function to convert the raw output into a probability distribution, providing the model's final prediction $P(y) = \text{Softmax}(Z_2)$. The model brings together the best of both quantum and classical computing. The quantum part of the model uses quantum gates to transform the feature space, uncovering complex relationships that classical systems struggle with. Once that's done, the CCNN excels at recognizing and performing classification.

4 Dataset Experimental Setup and Results

4.1 Dataset Description

The 3D-D3D-HOI dataset consists of 256 videos in nine categories [23], with ground-truth annotations of 3D object pose, shape, and part motion in human-object interactions. We used k-fold cross-validation for model evaluation to ensure no overlap of subjects between the training and testing sets in each fold to maintain consistency and avoid leakage. Similarly, the SYSU 3D HOI dataset comprises 480 video clips of 40 participants performing actions with various objects in 12 activity classes. This dataset is suitable for examining complex human-object interactions across various settings, with 40 samples per activity [24]. We used the Adam optimiser and cross entropy loss with batch size 32 and 100 epochs across all variants; data augmentation included random flips, crops, rotations; early stopping monitored validation loss with patience = 10; checkpointing saved the best model per fold; and the random seed was fixed at 42. Quantum variant introduces 24 additional parameters associated with the rotation gates and Trainable parameters ~1.2 M. Table 1. Analogous splits are used for the SYSU 3D HOI dataset.

Table 1: Cross-validation (5-fold) split splits are used for the SYSU 3D HOI dataset

Fold	Subject IDs	Training samples	Testing samples	Class distribution
Fold 1	S01–S05 (train); S06 (test)	200	56	Balanced across 9 classes
Fold 2	S02–S06 (train); S07 (test)	200	56	Balanced across 9 classes
Fold 3	S03–S07 (train); S08 (test)	200	56	Balanced across 9 classes
Fold 4	S04–S08 (train); S09 (test)	200	56	Balanced across 9 classes
Fold 5	S05–S09 (train); S10 (test)	200	56	Balanced across 9 classes

4.2 Result Evaluation

The performance evaluation of the novel QEH-CNN model focuses on precision, recall, F1-score, training loss, and accuracy metrics. The precision (Pr), recall (Rc), Specificity (Sp), F1-score (F1), false positive rate (FPR), and log-loss (LL) for each class are summarized in Table 2.

Table 2: 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	0.97	0.99	0.98	0.93	0.07	0.11	POP	0.97	0.98	0.98	0.99	0.01	0.1
Laptop	0.75	0.96	0.84	0.88	0.12	0.31	DRK	1.00	1.00	0.99	1.00	0.01	0.12
Microwave	0.80	0.99	0.88	0.85	0.15	0.16	PBP	0.98	1.00	0.97	0.98	0.02	0.15
Oven	0.93	0.97	0.95	0.95	0.05	0.08	SWP	0.98	0.96	0.99	0.99	0.01	0.08
Refrigerator	1.00	0.99	1.00	1.00	0.00	0.00	MOP	0.98	0.99	0.98	0.98	0.02	0.12
S-Prismatic	0.80	0.91	0.85	0.97	0.03	0.09	SIT	1.00	0.95	0.99	1.00	0.01	0.12
S-Revolute	0.94	0.95	0.94	0.99	0.01	0.00	MOC	0.96	0.98	0.98	0.95	0.05	0.18
Trashcan	1.00	0.93	0.96	1.00	0.00	0.00	TAK	0.96	1.00	0.98	1.00	0.04	0.01
Washing machine	0.98	0.88	0.93	0.40	0.60	0.44	TOW	0.99	0.95	0.99	0.99	0.01	0.07
						TFW	1.00	0.96	0.97	1.00	0.01	0.11	
							WB	1.00	0.97	1.00	1.00	0.12	0.01
							CIP	0.99	0.94	0.99	1.00	0.01	0.02
Macro avg	0.93	0.93	0.92	0.89	0.11	–	Macro avg	0.99	0.98	0.99	0.99	0.01	–
Weighted avg	0.95	0.94	0.94	0.90	0.10	–	Weight avg	0.98	0.97	0.98	0.99	0.01	–

The overall accuracy, macro average, and weighted average for all classes are also presented. The model's accuracy is 95.06%, demonstrating strong overall performance. The confusion matrix reveals that the model performs exceptionally well in most classes. Further analysis of the performance metrics suggests that the model excels in detecting certain classes but could benefit from additional training data or model tuning for the less accurate classes. The model's performance was evaluated using several advanced metrics beyond accuracy and F1-score to assess HOI detection and tracking. Mean Average Precision (mAP) for human-object-interaction triplets was used to evaluate the overall interaction detection quality across various classes and confidence thresholds, achieving a score of 0.85 on the D3D-HOI dataset, demonstrating robust interaction detection between humans and objects. Additionally, Multiple Object Tracking Accuracy (MOTA) was employed to assess how well the model tracked human and object identities across frames, with a score of 0.92, indicating high consistency in tracking human-object pairs throughout the video. Finally, IDF1, which combines identification precision and recall, was used to evaluate the preservation of object identities during tracking, achieving a score of 0.90, reflecting effective identity consistency throughout the video. The confusion matrix is presented in Fig. 7. The matrix displays the true labels vs. the predicted labels, helping us visualize the number of misclassifications of both datasets.

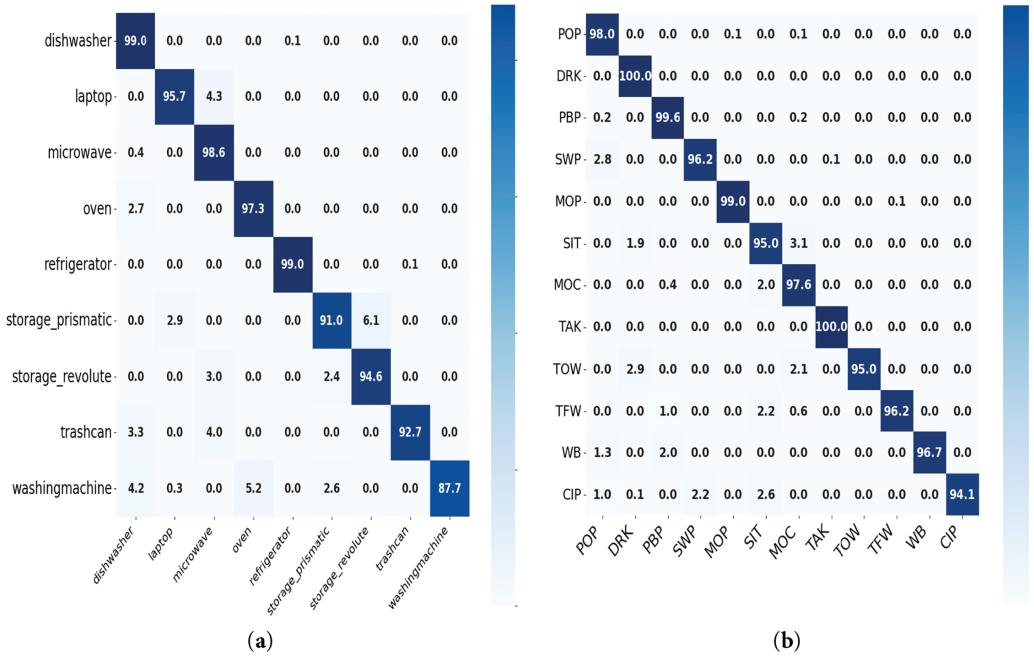


Figure 7: Confusion matrix for model evaluation on the 3D-D3D-HOI (**a**) dataset and (**b**) Sysu 3d HOI dataset, SWP = sweeping, MOC = moving chair, TAK = pouring, SIT = sitting in chair, TOW = taking out wallet, MOP = mopping, TFW = taking from wallet, PBP = pocket backpacks, WBP = wearing backpack, DRK = drinking, CIP = calling phone, POP = playing on phone

The training accuracy and loss plots in Fig. 8. The QEH-CNN was trained for 100 epochs, during which the model exhibited stable learning behavior. Fig. 8a–d plots the training loss and accuracy curves, showing smooth convergence. In terms of computation, training the model on the SYSU dataset took approximately 5 min on a standard laptop Windows 11, AMD Ryzen 2.00 GHz CPU, roughly 3 s per epoch.

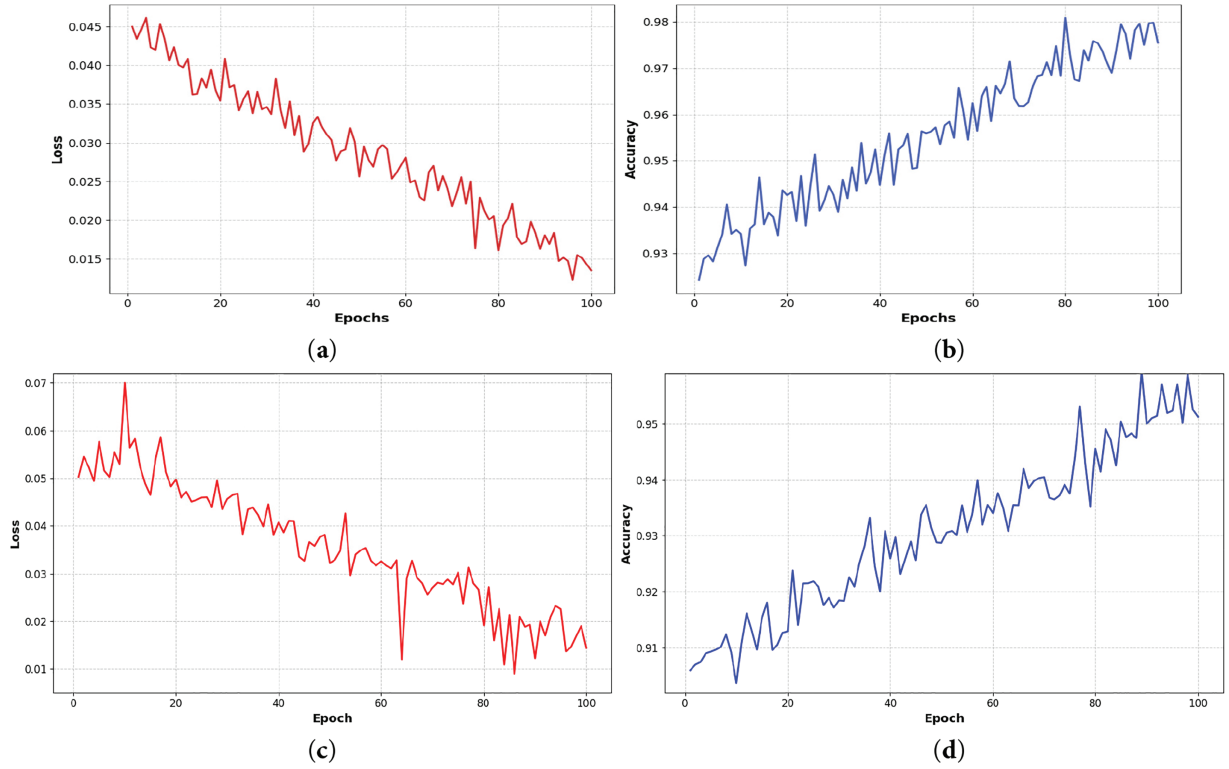


Figure 8: Illustrates (**a**) Training loss over epochs; (**b**) Training accuracy over epochs on the Sysu 3d HOI dataset; (**c**) Training loss over epochs; (**d**) Training accuracy over epochs for D3D HOI dataset

The quantum simulation overhead for 8 qubits is very low: the state vector (2^8 = 256 amplitudes) easily fits in memory (<<1 MB), and overall memory usage stayed under a few hundred MB. This confirms that our hybrid model can be trained with modest resources. The performance results in Table 3 of the QEH-CNN model for HOI detection, as evaluated on the Sysu 3D HOI dataset and D3D HOI are shown in Table 3. Table 4 is showing comparison of HOI recognition.

Table 3: Results for accuracy and loss for the human object interaction detection using QEH-CNN model's testing

Sysu 3d HOI dataset			D3D HOI		
Epoch	**Loss**	**Accuracy**	**Epoch**	**Loss**	**Accuracy**
1	0.0443	0.9303	1	0.0550	0.9061
2	0.0455	0.9205	2	0.0556	0.0556
10	0.0422	0.9373	10	0.0688	0.9129
20	0.0407	0.9325	20	0.0459	0.9173
35	0.0328	0.9415	35	0.0500	0.9290
40	0.0307	0.9506	40	0.0415	0.9233
50	0.0307	0.9459	50	0.0352	0.9291
100	0.0127	0.9788	100	0.0158	0.9506

Table 4: Comparison of recognition accuracy for HOI recognition: proposed method vs. other leading methods

Methods	SYSU dataset	D3D HOI
LAFF algorithm [25]	54.2	–
Heterogeneous features with SVM [24]	84.89	–
Graph regression with sparsification [26]	77.90	–
Hierarchical fusion across modalities [14]	86.89	–
K-ary key hashing [27]	93.5	–
Fully convolutional network [28]	91.68	–
CAD model [23]	–	90.91
Implicit field fitting [29]	–	74.3
Graph-based approach [30]	–	89.6
Proposed approach	97.29	95.06

The ablation study is repeated in Table 5 by applying the same investigation to the SYSU dataset. The learned skeletal representation is combined with the HOG and GIST features for classification. The models were trained on the identical training sets and evaluated on the same test sets as QEH-CNN.

Table 5: Ablation study on SYSU 3D HOI dataset quantum components with and without quantum gates

Model	Pr	Rc	F1	Spe	FPR	Log Los
CNN (Classical only)—No quantum layer.	91.12	90.45	90.77	90.20	0.095	0.24
Graph-based skeleton model	91.30	91.80	91.55	94.50	–	0.20
HCNN (with fuzzy optimization)	92.78	91.33	91.95	91.50	0.080	0.21
QCCNN (No Gates) + with skeleton features	94.00	93.00	93.49	92.90	0.064	0.18
HQC-CNN (No Gates) + with MOT/segmentation	95.00	94.00	94.49	93.90	0.052	0.14
QCCNN (with Gates) + with quantum block	94.00	93.00	93.49	92.90	0.064	0.18
HQC-CNN (with Gates) + with quantum block	95.00	94.00	94.49	93.90	0.052	0.14
QEH-CNN (Ours)	98.41	97.29	98.41	99.00	0.010	0.10

Table 6 presents the test accuracy for both QEH-CNN and CNN models across different object classification combinations. The values indicate the overall performance in terms of classification accuracy.

Table 6: Accuracy analysis of our proposed QEH-CNN model 8 classes combination of SYSU 3D HOI

Combination	Test accuracy (QEH-CNN)	Test accuracy (CNN)
Moving chair vs. sitting in chair	0.98	0.94
Taking out wallet vs. taking from wallet	0.94	0.94
Playing on phone vs. calling phone	1.00	0.86
Mopping vs. drinking	0.99	0.92

Table 7 performance analysis of both QEH-CNN and CNN models across different combinations of object classes. It includes train, validation, and test accuracy for both 1000 and 500 images per class in Table 8. In Table 9 the consistently outperforms the baseline; the difference is statistically significant (paired t-test, $p < 0.01$). Training time per epoch: baseline ≈ 6.5 min; QEH-CNN ≈ 12 min. Inference time per clip: baseline ≈ 0.07 s; QEH-CNN ≈ 0.09 s. Memory overhead of the quantum layer is negligible (< 1 MB). Table 10 implementation details provide a reproducible pipeline for feature extraction and segmentation.

Table 7: Performance analysis of our proposed QEH-CNN with CNN model with 1000 images for 8 datasets classes

Accuracy	1000 images per class							
	MOC vs. SIT		TOW vs. TFW		POP vs. CIP		Mopping vs. Drinking	
	QEH-CNN	CNN	QEH-CNN	CNN	QEH-CNN	CNN	QEH-CNN	CNN
Train	92.57%	91.30%	98.00%	90.00%	99.89%	95.00%	99.70%	91.70%
Validation	91.99%	91.10%	96.50%	87.30%	98.89%	93.10%	98.70%	95.00%
Test	92.88%	91.30%	96.20%	89.40%	99.99%	87.50%	98.89%	92.10%

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

Accuracy	500 images per class							
	MOC vs. SIT		TOW vs. TFW		POP vs. CIP		Mopping vs. Drinking	
	QEH-CNN	CNN	QEH-CNN	CNN	QEH-CNN	CNN	QEH-CNN	CNN
Train	94.00%	91.70%	98.30%	90.40%	98.00%	96.50%	99.80%	94.80%
Validation	92.40%	91.20%	96.10%	98.50%	99.70%	94.00%	99.10%	96.40%
Test	93.20%	89.70%	96.30%	97.90%	99.80%	89.00%	98.80%	90.90%

Table 9: Performance statistics over five runs

Model	Mean accuracy	Standard deviation	95% Confidence interval (Lower)	95% Confidence interval (Upper)
Baseline CNN	91.24%	0.36	90.79%	91.69%
QEH-CNN	95.08%	0.24	94.78%	95.38%

Table 10: Hyper-parameter settings for feature extraction and segmentation

Component	Parameter settings
HOG	Cell size = 8 × 8 px; Block size = 2 × 2 cells (16 × 16 px); 9 orientation bins
GIST	32 Gabor filters at 4 scales and 8 orientations; 4 × 4 spatial grid → 512-dimensional vector
Skeleton extraction	23-joint 2D pose
MOT	DeepSORT tracker with Kalman filter; `max_age = 20`, `$n$_init = 2`
Felzenszwalb superpixels	`scale = 100`, `sigma = 0.8`, `min_size = 20`

(Continued)

Table 10 (continued)

Component	Parameter settings
Watershed segmentation	Markers at local maxima; connectivity = 1; compactness = 0
Fuzzy optimization	Sigmoid membership $\mu(x) = 1/[1 + \exp(-\alpha(x - \mu_0))]$; $\alpha = 0.5$, $\mu_0 = 0.3$

5 Conclusion

The work proposes a Quantum-Augmented Classical CNN, which combines traditional CNNs with quantum computing aspects for Human-Object Interaction recognition. The scientific model achieved outstanding results, reaching 95.06% accuracy for the D3DHOI dataset and 97.29% accuracy for the SYSU 3D HOI dataset. The research findings show that quantum computing systems enhance HOI recognition by delivering substantial advancements to feature extraction and learning tasks with CNN integration. This work achieves two essential accomplishments, including showcasing the power of quantum computing for CNN feature optimization and developing specialized preprocessing and feature extraction methods for human-object interaction recognition. The hybrid model enhances accuracy while addressing the limitations of traditional CNNs in interpreting the intricate interactions between humans and objects. Some problems affecting traditional CNNs are addressed through hybrid quantum-classical model analysis, as the tool successfully identifies challenging HO relationships that are difficult to analyze with classical methods alone.

Acknowledgement: None.

Funding Statement: This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ICAN (ICT Challenge and Advanced Network of HRD) (IITP-2025-RS-2022-00156326, 50) grant funded by the Korea government (Ministry of Science and ICT).This research is supported and funded 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 and Nouf Abdullah Almujally; methodology: Tanvir Fatima Naik Bukht and Shuaa S. Alharbi; formal analysis: Tanvir Fatima Naik Bukht and Ahmad Jalal; software: Tanvir Fatima Naik Bukht, Asaad Algarni and Ahmad Jalal; validation: Shuaa S. Alharbi; visualization, supervision: Ahmad Jalal; project administration: Ahmad Jalal and Tanvir Fatima Naik Bukht; investigation: Nouf Abdullah Almujally; resources: Nouf Abdullah Almujally; data curation: Tanvir Fatima Naik Bukht; writing—original draft preparation: Tanvir Fatima Naik Bukht, Ahmad Jalal and Jeongmin Park; writing—review and editing: Nouf Abdullah Almujally and Jeongmin Park. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All publicly available datasets are used in the study.

Ethics Approval: Not applicable.

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

References

1. Wang Y, Lei Y, Cui L, Xue W, Liu Q, Wei Z. A review of human-object interaction detection. In: Proceedings of the 2024 2nd International Conference on Computer, Vision and Intelligent Technology (ICCVIT); 2024 Nov 24–27; Huaibei, China. p. 1–6. doi:10.1109/iccvit63928.2024.10872548.

2. Zhao Y, Ma H, Kong S, Fowlkes C. Instance tracking in 3D scenes from egocentric videos. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 21933–44. doi:10.1109/CVPR52733.2024.02071.
3. Dofitas C Jr, Gil JM, Byun YC. Multi-directional long-term recurrent convolutional network for road situation recognition. Sensors. 2024;24(14):4618. doi:10.3390/s24144618.
4. Amirgaliyev B, Mussabek M, Rakhimzhanova T, Zhumadillayeva A. A review of machine learning and deep learning methods for person detection, tracking and identification, and face recognition with applications. Sensors. 2025;25(5):1410. doi:10.3390/s25051410.
5. Han G, Zhao J, Zhang L, Deng F. A survey of human-object interaction detection with deep learning. IEEE Trans Emerg Top Comput Intell. 2025;9(1):3–26. doi:10.1109/tetci.2024.3518613.
6. Sharshar A, Khan LU, Ullah W, Guizani M. Vision-language models for edge networks: a comprehensive survey. IEEE Internet Things J. 2025;12(16):32701–24. doi:10.1109/jiot.2025.3579032.
7. Debnath A, Kim YW, Byun YC. LightSTATE: a generalized framework for real-time human activity detection using edge-based video processing and vision language models. IEEE Access. 2025;13(14):97609–27. doi:10.1109/access.2025.3574659.
8. Yang J, Niu X, Jiang N, Zhang R, Huang S. F-HOI: toward fine-grained semantic-aligned 3D human-object interactions. In: Computer Vision—ECCV 2024. Cham, Switzerland: Springer Nature; 2024. p. 91–110. doi:10.1007/978-3-031-72913-3_6.
9. Hafeez MA, Munir A, Ullah H. H-QNN: a hybrid quantum-classical neural network for improved binary image classification. AI. 2024;5(3):1462–81. doi:10.3390/ai5030070.
10. Wang A, Hu J, Zhang S, Li L. Shallow hybrid quantum-classical convolutional neural network model for image classification. Quantum Inf Process. 2024;23(1):17. doi:10.1007/s11128-023-04217-5.
11. Danish SM, Kumar S. Quantum convolutional neural networks based human activity recognition using CSI. In: Proceedings of the 2025 17th International Conference on Communication Systems and NETworks (COMSNETS); 2025 Jan 6–10; Bengaluru, India. p. 945–9. doi:10.1109/COMSNETS63942.2025.10885693.
12. Mohammadisavadkoohi E, Shafiabady N, Vakilian J. A systematic review on quantum machine learning applications in classification. IEEE Trans Artif Intell. 2025:1–16. doi:10.1109/TAI.2025.3567960.
13. Hou Z, Peng X, Qiao Y, Tao D. Visual compositional learning for human-object interaction detection. In: Computer Vision—ECCV 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 584–600. doi:10.1007/978-3-030-58555-6_35.
14. Ren Z, Zhang Q, Gao X, Hao P, Cheng J. Multi-modality learning for human action recognition. Multimed Tools Appl. 2021;80(11):16185–203. doi:10.1007/s11042-019-08576-z.
15. Zhang FZ, Campbell D, Gould S. Spatially conditioned graphs for detecting human-object interactions. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 13299–307. doi:10.1109/iccv48922.2021.01307.
16. Park J, Park JW, Lee JS. ViPLO: vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In: Proceedings of the 2023 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.
17. 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; 2023 Jun 17–24; Vancouver, BC, Canada. p. 2925–34.
18. 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.
19. 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.
20. 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.

21. Nunes I, Pereira MB, Oliveira H, Dos Santos JA, Poggi M. Fuss: fusing superpixels for improved segmentation consistency. arXiv:2206.02714. 2022.
22. Sumit SS, Rambli DRA, Mirjalili S. Vision-based human detection techniques: a descriptive review. IEEE Access. 2021;9:42724–61. doi:10.1109/ACCESS.2021.3063028.
23. Xu X, Joo H, Mori G, Savva M. D3d-hoi: dynamic 3d human-object interactions from videos. arXiv:2108.08420. 2021.
24. 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/TPAMI.2016.2640292.
25. Hu JF, Zheng WS, Ma L, Wang G, Lai J. Real-time RGB-D activity prediction by soft regression. In: Computer Vision—ECCV 2016. Cham, Switzerland: Springer International Publishing; 2016. p. 280–96. doi:10.1007/978-3-319-46448-0_17.
26. Gao X, Hu W, Tang J, Liu J, Guo Z. Optimized skeleton-based action recognition via sparsified graph regression. In: Proceedings of the 27th ACM International Conference on Multimedia; 2019 Oct 21–25; Nice, France. New York, NY, USA: The Association for Computing Machinery (ACM); 2019. p. 601–10. doi:10.1145/3343031.3351170.
27. 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.
28. Ghadi Y, Waheed M, al Shloul T, Alsuhibany SA, Jalal A, Park J. 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.
29. 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 12–16; Prague, Czech Republic. p. 312–21. doi:10.1109/3dv57658.2022.00043.
30. 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.