Multimodal Human Interaction Recognition Framework Using Multi-Features and Deep Learning Approach

Tanvir Fatima Naik Bukht
Faculty of Computing and AI
Air University, Pakistan
211893@students.au.edu.pk

Haita Alhaston
Department of Cyber Security
Umm Al-Qura University, Saudi Arabia
haita.f.alhaston@uq.edu.sa

Noif Alshamrari
Department of Cyber Security
POSTECH University, South Korea
n.alshamrari@pu.edu.kr

Nouf Abdullah Almujally
Department of Information Systems, Princess Nourah bint Abdulrahman University, Saudi Arabia
naalmujally@pnu.edu.sa

Ahmad Jalal
Department of Computer Science and Engineering
Korea University, South Korea
ahmadjalal@mail.au.edu.pk

Abstract: *Human Interaction Recognition (HIR) is one of the most important research topics in computer vision and pattern recognition that deals with the identification of specific interactions in static images and has several challenges that are related to the lack of temporal data, feature extraction, variability in image conditions, and the requirement of more accurate and interpretable robust models. However, current approaches face difficulties in recognizing the static images potential for interaction recognition, which results in a lack of effective algorithms using these resources. Addressing these gaps could potentially lead to great strides in the field. This paper aims to fill this gap by presenting a new Convolutional Neural Network (CNN)-based deep learning framework for interaction recognition, which integrates multimodal data for enhanced performance. The following steps are followed in the methodology: Preprocessing the images using Hue Saturation Value (HSV) color transformation to improve the image quality and silhouette extraction using Multiple Object Tracking (MOT) and Visual Background Subtractor (ViBe) techniques. We employ two distinct feature extraction approaches: Texton map for full body features and geometric attributes for skeleton features. The extracted features are then efficiently discriminated using Quadratic Discriminant Analysis (QDA). The analysis of our proposed framework suggests that the recognition rate on the Shakefive2 dataset is 90.2%, and the accuracy on the University of Lincoln (UoL) dataset is 92.3%. These results were compared to baseline models, such as traditional methods (e.g., handcrafted features), showing improved performance. These results show that the proposed method is a good solution for human interaction recognition based on static images. This research helps to enhance state-of-the-art deep learning-based algorithms for human interaction recognition that could be used for human-computer interaction, video analysis, and surveillance, and thus contributes to the field of computer vision.*

Keywords: *Image analysis, pattern analytics, body pose, interaction classification, deep learning.*

Received March 9, 2025; accepted July 31, 2025
https://doi.org/10.34028/iajit/23/1/12

1. Introduction

Human Interaction Recognition (HIR) is an important field of research in computer vision and pattern recognition concerned with the recognition of particular human interactions based on visual information. HIR has a lot of potential but it has a lot of challenges especially when it comes to static images which do not have the time aspect of video data. The main problems are the complexity of feature extraction, the inconsistency of image conditions (e.g., lighting and background clutter), and the necessity of more accurate and interpretable models that can address these issues. The existing methods of HIR are mostly based on hand-crafted features that have limited capabilities to reflect the complex nature of human interactions. Such models also tend to underutilize the potential of static images, and it is hard to create robust algorithms that can recognize a large variety of interactions.

To address these gaps, this paper proposes a Convolutional Neural Network (CNN) based deep learning framework that integrates multimodal data, offering a more effective solution for recognizing human interactions in static images. The framework leverages image features, geometric attributes from skeletons, and potentially depth information for more robust recognition. Although the current work primarily focuses on static images, the proposed framework is designed to be extensible to other sensor-based modalities like depth or audio data, improving the interaction recognition process.

The framework uses the latest methods including Hue Saturation Value (HSV) color conversion, silhouette extraction with Multiple Object Tracking (MOT) and Visual Background subtractor (ViBe), and feature extraction with Texton maps and skeleton features. Using these techniques, we hope to enhance the

performance of recognition and help to develop HIR in practice. The CNN is the main classifier within our framework, which learns and classifies the interactions using the multimodal features extracted. Quadratic Discriminant Analysis (QDA) is applied as a feature discrimination step to improve the performance of the CNN by effectively discriminating the extracted features prior to final classification.

In the considered context, the CNN will represent the main classification model, which will exploit multimodal characteristics to distinguish between human interactions. QDA is used as a feature discrimination step to enhance the performance of the CNN by efficiently separating the extracted features before final classification. HIR can be used in a variety of fields such as surveillance, sports analytics, human-computer interaction. In surveillance, accurate human interactions are detected to add security systems by identifying suspicious actions. In surveillance, accurate recognition of human interactions can enhance security systems by identifying suspicious behaviors. In sports, HIR can be used to analyze player movements and interactions, improving testing and training methods. Additionally, human-computer interaction benefits from HIR by enabling more intuitive and natural user interfaces, where machines can understand and respond to human gestures and actions [2].

HIR is used in many areas, including sports, healthcare, security, and human-computer interfaces. The potential to precisely calculate and interpret human activities profoundly affects human conditions and the effectiveness of various applications. One promising strategy is implementing CNN [1]. Such models consider sequential dependencies to model the features of complex interaction patterns, delivering reliable recognition results. Filtering and identifying relevant features from various activities performed under different circumstances is challenging. Some are primitive and may not capture all aspects of human motion, while others like deep learning may need large amounts of labelled data for training [37]. Lighting, background clutter, and occlusions affect recognition accuracy when lighting and background change or there are occlusions [3]. Most of the existing systems fail to operate optimally under such conditions. It was found that many methods fail to properly use temporal information which is essential for discriminating between similar activities. This is quite a limiting factor, especially in dynamic environments, because it can lead to misclassification [5].

This paper contributes to the field of computer vision by presenting a novel CNN-based approach for HIR, offering significant improvements in recognition accuracy and robustness. Our proposed framework not only enhances the state-of-the-art algorithms but also demonstrates the potential of deep learning to tackle longstanding challenges in the recognition of human interactions from static images. The proposed system comprises the following key contributions:

1. HSV transformation: improves the quality of image frames by providing better color differentiation that helps in feature identification.
2. Silhouette extraction: uses MOT and ViBe techniques to obtain a precise estimation of shapes of human body.
3. Feature extraction: Texton maps is used for the full-body and geometric characteristics of skeleton features to better understand interactions.
4. QDA: exclusively discriminates features from the extraction process, enhancing the classification rate when fed to the CNN.

This paper is organized as follows: 1st Section summarizes the previous studies and the current state-of-the-art in HAR to pinpoint the existing methods, their limits, and the gaps in knowledge. The 2nd Section of our paper discloses the proposed framework and details of CNN, which offers full body textures and geometric features. 3rd Section provides the experimental setup, including the datasets, metrics, and implementation details. The results and analysis from the experiment are displayed in 5th Section conclusion, and future research perspectives in the field of HAR are provided.

2. Related Work

HIR has gained enormous recognition in multiple domains over the past years, such as healthcare, sports, surveillance, and human-computer interaction. Numerous proposed approaches target this challenge, using different methods such as Machine Learning (ML), computer vision, and sensor-based systems. This part discusses the methods used widely and explains their advantages and disadvantages.

2.1. Human Behavior Interaction with Machine Learning

Human behavior interaction with ML deployment of this surveillance [21] and suspicious interaction [4] detection methods also form an essential aspect of HBI applications as they help determine deviant behavior in the commons, identify threats and protect people. HBI also facilitates tracking of patient mobility, observing the performance of the patients in exercises and physics therapy, and even modifying the treatment regimens [6] created a Support Vector Machine (SVM), smartphone, real-time interaction identifying framework which was 87% accurate. The interaction recognition system described in their study uses ML methods on depth camera skeleton data to improve the instrument's dependability. Multiclass SVM and X-mean algorithms are used to categorize interactions using this tool based on postures. The method is better than the best art techniques that can process input data in no more than 4 seconds. Another researcher was building dynamic texture descriptors for human interaction detection,

which they noted could be used to simplify the computation. This is about picture data and contrasting outcomes with the best strategies using the progression of computer vision research [7].

HAR has widely adopted ML techniques that extract discriminative features and train interaction classifiers [8]. They proposed a multi-layered framework that combines deep neural networks with Long Short Term Memory (LSTM) units to learn temporal dynamics of activities. They showed that their results achieved improved recognition accuracy over conventional ML approaches. Similarly, Gemeren *et al*. [11] use SVMs to learn discriminative hyperplanes embedded with various high-dimensional feature spaces to support interaction recognition. In spite of that, a number of ML approaches still depend on some hand-made features, and it is hard to reflect the complexity of human activities with them.

2.2. Human Behavior Interaction Deep Learning

In human behavior interaction with deep learning [10] a fuzzy deep learning algorithm is presented to evaluate users of the lower limbs exoskeleton's daily activities based on real-time walking data, accomplished transition of gait mode and dynamic dataset. But nowadays, the features that are used for the recognition of human movement are limited, such as the joint positions of skeletal joints [12] or motion trajectories [13], these methods typically have a high degree of accuracy, but they may not be able to capture all the discriminating information present in human behavior. To cope with this drawback, our suggested method aims to intensify HAR that are equipped with full-body texture and geometric features. Full-body texture features are quite smooth and fine grain surface details such as clothing patterns and skin texture can support interaction recognition. Geometric characters include the space relationships between body parts, which implies depicting the structure's characteristics and the body's configuration during the activities.

Compared with other kinds of Neural Network (NN) architectures, the CNNs are overpowered. The reason is that CNNs can learn richer, higher-order features and that input images have a deep pixel correlation [15]. As for the image classification, when Deep Convolutional Neural Networks (DCNN) were successfully applied, object detection also progressed considerably with deep learning methods. Being convolutional neural networks, DCNNs inherently generate hierarchical features to map raw pixel values into semantic features, learn automatically from training data, and are proficient in discriminant performance in intricate circumstances. This, in turn, resulted in object detection algorithms using deep convolutional neural networks with end-to-end optimization and richer features representation [16].

Table 1. Related work for existing human interaction techniques and recognition model.

State-of-the-art models	Main contributions	Limitations	Proposed model comparison
Handcrafted features [22]	High interpretability, less computational cost. reaching overall accuracies of 0.87 and 0.88.	Limited adaptability to complex interactions. Struggles with dynamic conditions (lighting, occlusion).	Our model overcomes these limitations by using multimodal features, including Texton maps for full-body textures and skeleton features to capture more nuanced interactions.
ST-GC N [23]	Good at capturing spatial and temporal relationships.	Does not fully capture detailed motion or subtle cues between body parts.	We improve robustness and accuracy by combining skeleton features with image-based features (Texton maps), providing a richer representation.
Two-stream network [9]	They developed a two-stream network that combines spatial and temporal information for skeleton based action recognition. results show that with a percentage of 80 %.	Requires extensive training data for optimal performance.	Our model achieves similar performance with reduced complexity, using fewer training samples and incorporating multimodal data for better generalization.
RNNs with CNNs [24]	Captures long-term dependencies for action recognition. The accuracy of each of the competing methods is above 90%.	Struggles with real-time processing and computational load.	Thus, we rely on skeleton features that encode motion temporal aspects and make use of CNNs.
Interaction transformer [25]	A Transformer based model for human interaction recognition that captures long range dependencies and complex interactions. Accuracy of the proposed method for body partdetection is 90.91%.	It suffer from optimal parameter tuning and model selection for best performance.	The power of CNNs for feature extraction is leveraged in our model, which can be potentially combined with attention mechanisms to further improve interaction modeling.
Optical flow [14]	Effective for capturing motion in controlled conditions. Proposed method accuracy 87.2%.	Sensitive to noise and occlusions.	Our framework is more robust, addressing noise and occlusion issues through advanced preprocessing (HSV, MOT, ViBe).
Hybrid approaches [26]	Combines strengths of multiple methods. IGFormer achieves state-of-the-art performance 85.4%.	Increased complexity and computational load.	One can consider our model as a hybrid by combining the skeleton features with Texton maps which are image features.
Multi-modal approaches [27]	Utilizes diverse data sources. Achieve 76.7% accuracy with convnet +lstm+RGB.	Integration challenges and data alignment.	We proposed an improved model which can handle multi-modal data sources for better results.

CNNs and RNNs are very much suggested as possible networks for face recognition, which is accomplished with recent deep learning studies [17]. These methods can acquire a structural description of images and videos articulating activities [18]. The current CNN-based approaches are excellent data domain such as medical image classification or object detection, However, they need a massive number of labeled data as input and the computational cost is high. Other human behavior-crafted features like Motion

History Image (MHI), optical flow, and 3D Convolutional Neural Networks (3D CNNs) are now used to determine Human. Furthermore, different research studies have proposed different techniques to improve the success of human pose detection and recognition, such as combining CNN with Hidden Markov Model (HMM) [19]. Then, in this setup we use the CNN to extract features from the image and HMM as a model for the temporal information of the activities. Human interaction recognition is a challenging but promising field in computer vision, where the systems learn to recognize human interactions in videos. They have been used in areas such as monitoring security cameras or enhancing the ways that humans interface with machines. Although there have been some breakthroughs on this front, challenges still need to be addressed to develop accurate and generalizable interaction recognition models [20]. Table 1 is representing related work summery of human interaction recognition model.

3. Methods

In this work, we describe a CNN-based interaction recognition technique. We suggest 1st with the HSV color transformation as a preprocessing step that will help increase the frame clarity. Then the MOT and ViBe steps are applied to the template to get the silhouette. Feature extraction is performed using two distinct approaches: Texton map is employed to obtain all body traits, and the point features of geometric visualization gain skeleton features. QDA effectively discriminates the released features. Feature extraction and discrimination and then apply CNN. System architecture shown in Figure 1.

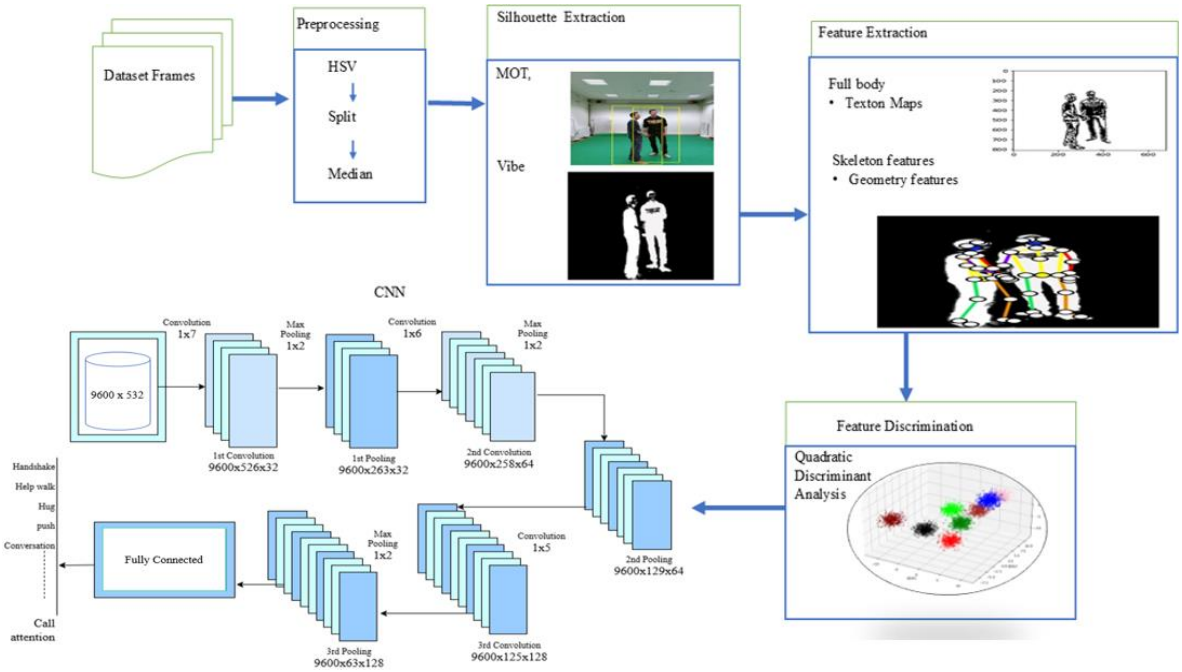


Figure 1. Structural diagram of our novel proposed model.

3.1. Preprocessing: Adaptive Mean Filter

First step in preprocessing is essential for removing noise and extracting features from frames. With this step, we can accurately predict human activities. This paper proposed a preprocessing technique to address this problem, which is one of the concerns. The method consists of two steps:

a) Converting the color space into HSV.
b) Choosing the most suitable channel and then smoothing the image with the median filter.

Figure 2 illustrates the outcome of such a process. Transformation of an HSV color space maximizes the difference between the pixels of the input video frame. This transformation employs the original video frame denoted as $I(p,q)$ as its operands. Set up $R(p,q)$, $G(p,q)$, and $B(p, q)$ to be the red, green, and blue channels of the image that needs to be processed. The HSV channels are computed as follows:

$$V = M\left(M(R(p,q)), M(G(p,q)), M(B(p,q))\right) \quad (1)$$

$$S = \begin{cases} \dfrac{V - \min[R(p,q), G(p,q), B(p,q)]\{V\}}{V}, & if\ V \neq 0 \\ otherwise \end{cases} \quad (2)$$

$$H = \begin{cases} \dfrac{60(G(p,q) - B(p,q))}{V - min(R(p,q), G(p,q), B(p,q))} \\ where\ V = R(p,q),\ and \\ 120 + \dfrac{60(B(P,Q) - R(P,Q))}{V - min(R(P,Q), G(P,Q), B(P,Q))}, \\ if\ V = G(P,Q)\ (3) \\ 240 + \dfrac{\{60(R(p,q) - G(p,q))\}}{\{(V - min(R(p,q), (G(p,q), B(p,q)))\}}, \\ \{if\}\ V = B(p,q) \end{cases} \quad (3)$$

Here, M represents the number of training samples used in the model, and $R(p, q)$ denotes the color values at pixel

(p, q) in the image. The corresponding channels of the HSV image, which are *V*, *S*, and *H* for value, saturation, and hue, respectively, are put in this formula. The color hue is normalized in the range [0, 1] with 360 when being divided. Lastly, the enhanced contrast image is yielded by merging the hue, saturation, and value channels into an HSV image, followed by its conversion back to the BGR color space. To address the noise, a median filter is applied to the channel that shows the best performance with the goal of reducing the noise.

a) The original frame. b) The HSV channel. c) The split channel, and the median filter applied.

Figure 2. HSV transformation is used to improve the frame.

HSV transformation is used to improve the frame in Figure 2-a) the original frame, in Figure 2-b) the HSV channel, in Figure 2-c) the split channel, and the median filter applied.

In other words, the preprocessing method used in this study increases the contrast in video frames by converting them to the HSV colour space. The best channel is chosen, and a median filter is thus applied to remove unwanted noise. This method requires more accuracy in human activations.

3.2. Silhouette Extraction

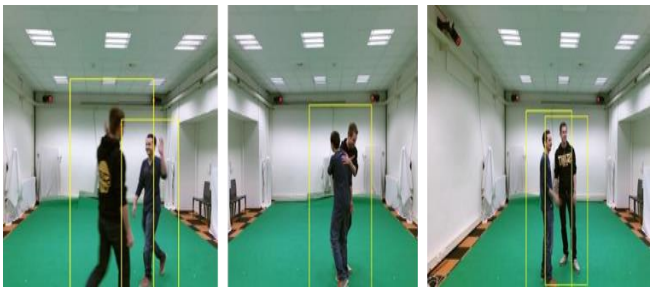
Silhouette extraction is a crucial task for feature extraction, used to capture accurate human shapes and movements [28, 29]. Here, we are speaking about how to have precise shapes of people. The critical task is to get the silhouette in place to extract and detect the best features. We are using two techniques for efficient extraction of silhouette 1st multi object tracking which help to track object then apply ViBe for batter silhouette extraction.

3.2.1. Multiple Object Tracking (MOT)

MOT is a non-overlapping data processing technique that tracks targets simultaneously from one frame to the next during a given time. MOT is widely implemented in security and self-driving vehicle analysis [30]. Primarily, the function of the MOT-based technique is to accurately determine the location of the objects and their identity while at the same time predicting their upcoming state [31]. Such regularization is associated with challenges, such as when obstacles or objects are often processed. In recognition of the need to maneuver through these obstacles, MOT algorithms deploy different procedures, including data association, motion prediction and object representation. An alternative approach is the implementation of object tracking technologies which consist of detecting objects in every frame of video frame data to be processed and applying a group matching technique that associates these detections during consecutive frames thus forming tracks. Most of MOT techniques are evaluated using measures that consider the numbers of true positives, false alarms and missed detections, such as tracking accuracy, precision, and recall.

$$P\left(Track|Detect\right) = \frac{P\sum_n^{Detect} \left|\sum_n^{Track} \cdot P\sum_n^{Track}\right.}{P\sum_n^{Detect}} \quad (4)$$

The probability of a track given a detection in *P* (*Track*|*Detection*) is obtained by multiplying the probability of the detection given the track the probability of the track itself, and the probability of the detection across all tracks. This equation is frequently used in the data association methods for MOT also results shown in Figure 3.



a) High five. b) Hug. c) Explain.

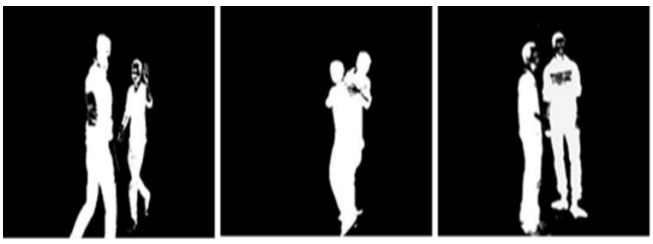
Figure 3. The figure illustrates silhouette extraction results of MOT.

3.2.2. Visual Background Subtractor (ViBe)

ViBe is an algorithm employed for background subtraction, a core procedure in many responsibilities of computer vision such as object detection and tracking. The background removal process is a technique that isolates the objects present in the foreground from the background in an image sequence [32]. The ViBe algorithm has become well known for its backdrop subtracting algorithm owing to its simplicity and efficiency. It represents the background by keeping the same set of statistics pixel-wise distribution distributions. Such covers are real-time spots of background changes. Through measurement of the new frame with the previous background model, ViBe validates the accuracy of the background and updates the respective distribution. The algorithm determines if the pixel belongs to the foreground or background area through pre-established thresholds and statistical measures. The ViBe algorithm has excelled in situations that fall into a category of difficult complexities like dynamic backgrounds, illumination changes, and camera motion.

$$B_t(x) = \frac{I_{(CV-BPV)}}{B_{t-1}} \therefore (x) > \emptyset B_{t-1}(x) \quad (5)$$

The term, $B_t(x)$, in Equation (5) refers to the background (or background pixel value) at position x in frame t. It (x) denotes the pixel value of the current image capture at location x The dissimilarity function $d(x)$ is used as a metric to quantify the dissimilarity between It (x) and the background model. If $d(x)$ goes over threshold θ, the pixel is taken as the foreground; otherwise, it is taken as the background pixel. Current pixel value as *CV* and background model pixel value as *BPV*. This formula is used in the ViBe algorithm to update the background model, result shown in Figure 4.



a) High five. b) Hug. c) Explain.

Figure 4. The figure illustrates silhouette extraction results of ViBe.

3.3. Feature Extraction

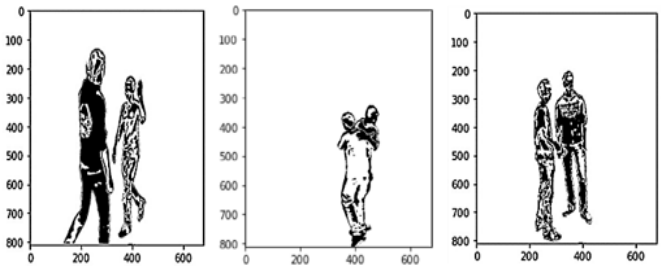
We employed Texton maps for full-body feature extraction, enabling effective representation and characterization of visual and textual elements in the data.

3.3.1. Full Body Features

Texton maps segment textures in images to aid in object recognition and scene understanding. Texton maps are used to segment an image into different textures so that the computer can analyze different Textons. This technique is very useful in object recognition, scene understanding, and the generation of realistic textures. Texton maps represent another important instrument for assessing an image's structural and compositional attributes by analyzing its texture information. Thus, this is becoming more precise and detailed in examining and categorizing the objects depicted in the image. Texton maps can supplement image segmentation algorithms that might use the image texture information in a more generalized form. The patterns of the Texture are classified into Textons.

$$T(x,y) = \sum_{(t)\in N} \min_{(u,v)\in N} d\big(I(x,y), T(t,u,v)\big) \qquad (6)$$

Where $T(x, y)$ Texton map value at the pixel coordinates (x, y) is calculated by Texton of the most suitable one (t) which is in the nearby region (N). Through function d (), we find the distance between the strength of the picture at coordinates (x, y) and the Texton value at (t, u, v), the latter being the coordinates of the neighborhood. The algorithm designates a pixel that is nearest to neighbors, and the final value of the Texton map, $T(x, y)$, is this distance. The Texton map, $T(x, y)$, will result when the argmin algorithm is performed to trace the Texton that achieves the smallest distance and results shown in Figure 5.



a) High five. b) Hug. c) Explain.

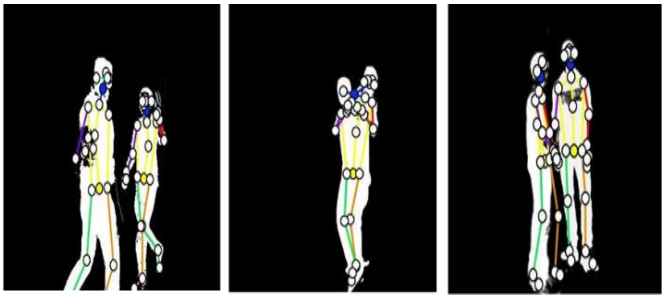
Figure 5. Illustrates feature extraction results of Texton maps.

3.3.2. Skeleton Features Extraction

The skeletal geometry features are very useful when the information is obtained from the skeletal structure to determine human motion [33]. These qualities include the positions and geometries of the important points of the human skeleton, such as junctions and bone length. The skeleton geometry features are often acquired using the method that is based on the distances between the joints of the skeleton, and the distances are measured using the euclidean distance metric. Calculations of distances of the joints in particular combinations provide hints regarding human poses and movements.

$$d_{lm} = \sqrt{(p-p_m)^2 + (q-q_m)^2 + (r_l - r_m)^2} \qquad (7)$$

Where d_{lm} is the euclidean distance between l^{-th} and m^{-th} joints of a skeleton. The coordinates of the l^{-th} joint is (pl, ql, zl) and the coordinates of the m^{-th} joint is (pm, qm, zm). The formula uses the square root of the square sum of differences between the p, q, and r parameters as the three-dimensional distance. The skeleton geometry features can be used to create Equation (7) motion capture systems, which produce a lot of information about the structure and movement of the human being. Results are shown in Figure 6 and Algorithm (1) extracts dynamic information of human skeletons in video frames, which includes joint displacement, statistical displacement and angles/orientation. These characteristics form a baseline set to understand human movement and interaction and serve as the basis of a follow-up process and machine-learning operation aimed at identifying specific types of interactions.



a) High five. b) Hug. c) Explain.

Figure 6. Skeleton geometry features result shown as High Five, Hug and Explain.

Algorithm 1: Skeleton joints keypoints and features extraction.

Input: Image frames
Output: Extracted joins features and key points
Initialize empty feature vectors for
-Joint displacement Dj
-Statistical displacement μ_j, σ_j^2
-Angles /orientation
for t = 1 to n do
If t > 1 then
for each joint j do
Calculate displacement vector Dj(t) between joint j in frames t-1 and t
Append Dj(t) to joint displacement feature vector
end for
end if
Store skeleton joints for statistical displacement and angles/orientation calculations
end for
Mean : $\mu_j = \frac{i}{n-1}\sum_{t=2}^{n}$ Dj(t)
Variance: $\sigma_j^2 = \frac{i}{n-1}\sum_{t=2}^{n}$ Dj(t) − $\mu_j 2$
for each joint j do
Append μ_j and σ_j^2 = to statistical displacement feature vector
for t = 1 to n do
for each joint j do
Determine joint angles and orientation relative to body segments
Append joint angles and orientation features to angles/orientation feature vector
end for
end for
Return: statistical displacement feature vector, joint displacement feature vector, and angles/orientation feature vector

The extracted features are then used for further classification and interpretation in subsequent stages.

$$C_y(a,b) = exp\left(-\frac{(a-a_y)^2 + (b-b_y)^2}{2\sigma_y^2}\right) \quad (8)$$

Where $C_y(a, b)$ is the confidence map for keypoint, (a_y, b_y) is the location of key point, σ_k the standard deviation controls the confidence spread around the key point location.

3.4. Feature Discrimination Analysis

QDA is a statistical classification technique that is being applied to determine the probability of an observation to belong to a particular class, or is it [34]. In this context, QDA is based on the principle that every class adheres to a multivariate normal distribution, and it estimates the parameters of the Gaussian process to forecast results. While Linear Discriminant Analysis (LDA) requires the classes to share the same covariance matrix, QDA provides the alternative possibility of fitting one class with a unique covariance matrix. This implies that QDA, through its methodologies, can detect any additional facets between the variables. The decision boundary in QDA is defined by a quadratic equation, which makes it a nonlinear classifier. The goal of QDA is to maximize the posterior probability of each class given the observed data. To achieve this, QDA calculates each class's discriminant function, represented by a quadratic equation. The discriminant function for class k is given by:

$$e^{-i\omega t} = ln(p(C_k)) - \frac{1}{2}ln(|\Sigma_k|) \quad (9)$$

$$\max_{0\le x\le 1} xe^{-x^2} = \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}d\,ln(2\pi) \quad (10)$$

$$g_k(x) = e^{-i\omega t} - \max_{0\le x\le 1} xe^{-x^2} \quad (11)$$

Where $gk(x)$ is the discriminant function for class k, in e^{-iwt} $p(Ck)$ is the prior probability of class k, Σk is the covariance matrix for class k, in $\max_{0\le x\le 1} xe^{-x^2}$, μk is the mean vector for class k, x is the input vector, and d is the dimensionality of the input space. Based on the discriminant functions, QDA assigns the observation to the class with the highest discriminant value. By considering the quadratic terms, QDA can capture more complex decision boundaries compared to linear classifiers like LDA. However, QDA requires more parameters to be estimated and may be more prone to overfitting when the number of training samples is limited. Figures 7 and 8 represent features fusion and discrimination result using QDA.

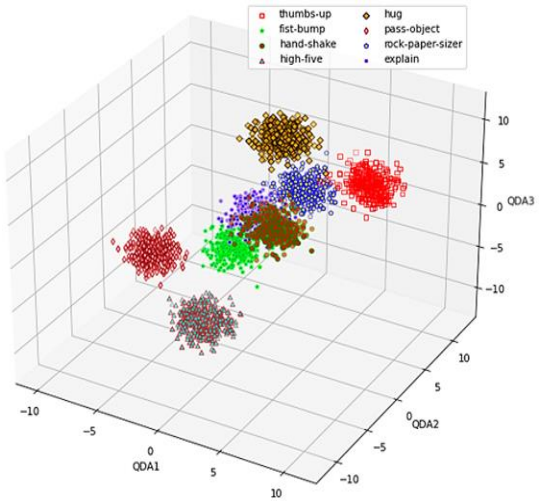


Figure 7. The features fusion and discrimination result using QDA on Shakefive2 dataset.

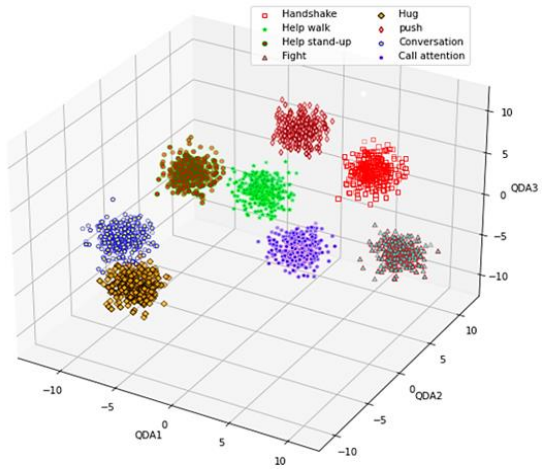


Figure 8. Features fusion and discrimination result using QDA on UoL dataset.

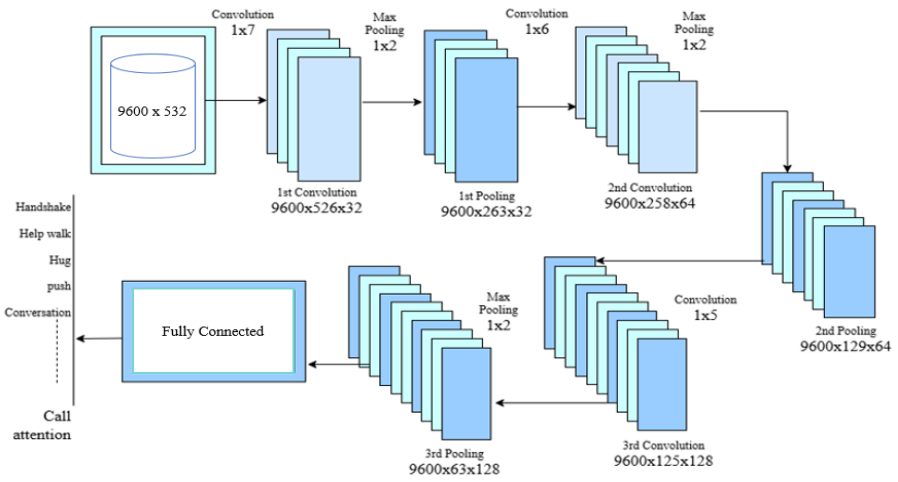


Figure 9. Standard architecture of CNN.

3.5. Convolutional Neural Network

Human interaction recognition is a typical and well-used task for CNN as they can automatically learn and classify different activities using input data. CNNs, then, can examine features which is already extracted. Those kinds of data capture such spatial and temporal patterns, which give chances to determine activities like hug, punch, push, or other specified activities. CNNs learn to recognize human activities with an incredible accuracy and can be used in real time systems for instance fitness trackers and monitoring systems in health care and sports analytics. Architecture of human interaction recognition shown in Figure 9. It has been implemented using a hybrid approach where convolutional neural network is used as a classifier on manually extracted visual features and raw images. In particular, Texton maps are computed in order to retain texture data, but geometric features based upon skeletons are used to represent spatial data. The resulting descriptors are concatenated and are used as inputs to CNN model. This process allows the CNN to leverage its powerful classification capabilities, not just on raw image data but also on the manually curated features, thereby enhancing the model's ability to recognize complex human interactions.

3.5.1. Convolutional Layers

The CNN architecture consists of three convolutional layers. Convolutional layers are the foundation of CNNs that control perceiving spatial hierarchies of features from input data. For HIR, the pre-extracted features are directly fed to Convolutional layers for classification. The 1x7 sized 32 filters are used in the first convolutional layer, which gives us an output feature map of size 9600x526x32. The calculation of this dimension takes into account the valid padding. The second convolutional layer uses 64 filters of size 1x6 and outputs 9600x258x64. The third convolutional layer uses 128 filters of size 1x5, giving us an output of 9600x125x128. We also want to mention that after each convolutional layer, we add activation functions ReLU and bias terms to improve model performance.

$$Conv^{\{(l+1)\}}(i,j) = ReLU(u) \quad (12)$$

The activation value of a neuron at position (i, j) of the feature map in a convolutional layer, after a convolutional layer in a CNN is denoted by $Conv^{\{(l+1)\}(i,j)}$. First we need to compute u, which is a weighted sum of the previous layer's inputs plus a bias term and then we just multiply that with a frame drop probability to compute this value. The *ReLU* activation function plays an important role in this process, bringing much needed non linearity to the network. Essentially, *ReLU* only looks at the input u and sets it to 0 if u is less than 0 or to u if u is nonnegative, so it only looks at positive values.

$$ReLU(u) = \sum_{\{a=1\}}^{\{v\}} \Omega\left[i, c, (a-1) + \frac{y+1}{2}\right] W^{\{a\}[i,c]} + k_{\{c\}}^{\{a\}} \quad (13)$$

The *ReLU* activation for a particular neuron in the CNN is calculated in detail by Equation (13). This involves a sum over a range of values most likely corresponding to different values of the previous channel (or feature) from which these vectors are derived. Accessing data from a multidimensional array or tensor, which is the inputs or the feature maps, is denoted by the notation $\Omega[i,\ c,\ (a\text{-}1)+\frac{y+1}{2}]$. $W^{\{a\}[i,c]}$ is the weight for the contribution of each feature or chanel, and $k_{\{c\}}^{\{a\}}$ is bias term that multiplies all output. This equation shows how the network processed different inputs and applied its learned parameters to generate meaningful activation values, so the network learns complicated patterns and makes predictions.

3.5.2. Pooling Layers

Pooling layers are mainly used to down-sample feature maps and generate information summaries. This helps reduce the complexity in the subsequent layers and hence saves computation. We apply the down-sample

max-pooling layer to each convolutional layer to reduce the size of feature maps. The first pooling layer has a 1x2 window doing a 1/2 of spatial reduction on the feature vector axis. The output is 9600x263x32. The second and third pooling layers also use 1x2 max-pooling, giving outputs of 9600x129x64 and 9600x63x128.

$$Pool^{\{l\}}(i,j) = \max\left(Conv^{\{l\}\left(i,((j-1)x(m+n))\right)}\right) \quad (14)$$

$Pool^{\{l\}}(i, j)$ is the result of pooling the feature map of layer l at position (i, j) taken in the context of a neural network. $Conv^{\{l\}(i,\ ((j-1)\text{x}(m+n)))}$ represents the area of the input feature map which is under examination by pooling window and is specific to that window. In this pooling operation, this area selects the maximum value, down sampling the feature map and reducing its size.

3.5.3. Fully Connected Layers

Through fully connected layers, the classification component becomes the central part of the CNN. These layers meet the previously extracted features' inputs and make decisions based on the learned representations. This enables the CNN to detect intricate links between the features and the interaction classes. The fully connected layers achieve the task by performing the matrix multiplications and nonlinear transformations, which, help to transform the pre-extracted features into class probabilities or scores used for the precise classification and recognition of human activities images.

$$FC^{\{(l+1)\}} = ReLU\left(\sum_k W_{lk}^0\, x_{lk}^0 + b_{\{l\}}^{\{a\}}\right) \quad (15)$$

Then, we move on to the fully connected layer, $FC\text{^}\{(l+1)\}$ is the output of a neuron in that layer. W_{lk}^0 represents weight between the k-th neuron on the previous layer and the *l-th* neuron on the current layer. The activation value in the neuron of the *k-th* layer in previous layer is symbolized x_{lk}^0. And $b_{\{l\}}^{\{a\}}$ is the term bias for the lth neuron in this layer. The output of a neuron is given as, we weight the activations from the previous layer getting the basic weighted summation of the biases on that level and finally applying *ReLU* activation on it. Then the process take place in which input is made to the neuron and output of the neuron is generated , so that the neuron can learn and predict.

The current study utilized TensorFlow version 2.4 to train a CNN model. The training parameters included a batch size of 32, a learning rate of 0.001, and the Adam optimizer that was applied to 50 epochs, where early stopping was applied in order to prevent overfitting. These decisions were made with care and a balance between performance capacity and computational performance with clear documentation so that it can be reproduced.

The datasets-Shakefive2 and University of Lincoln (UoL) were separated into 70 percent training, 15 percent validation and 15 percent testing. These ratios provided sufficient availability of training data along with close evaluation by the use of separate validation and test sets. As far as hyperparameters are concerned, a 3x3 median filter was added to remove noise in the input frames. The ViBe algorithm was tuned in the background update rate to 0.5, and the MOT algorithm was tuned to a track association threshold of 30 pixels and a maximum distance of 50 pixels between the successive frames to enable the tracking of the objects reliably.

4. Result and Analysis

In this paper, we implemented CNN as a classifier to evaluate the effectiveness of the presented approach. The experiment was carried out very carefully, with all the steps executed correctly, and the resulting numerical data was subjected to detailed scrutiny.

4.1. Dataset Description

4.1.1. ShakeFive2

ShakeFive2 focuses on dyadic human interaction in the dataset. The dataset comprises 8 different modes of Interaction: Fist bump, handshake, high five, hug, pass object, thumbs up, rock-paper-scissors, and explaining. with this dataset under our examination, our study aims to discover intricate connections and the general patterns among these human communications. Consequently, through our research, we anticipate adding to the knowledge base of human behavior and assisting in the development of intelligent systems development that facilitate social interaction interpretation and response to it.

4.1.2. UOL 3D Social Activity

UoL 3D social activity dataset data collection refers to two persons who are involved in social communication. The dataset comprises eight distinct social activities: greeting, hug, handshake, help stand up, help walk, push, conversation, fight, and call attention. Sessions were generally recorded in interaction samples, each about 40 to 60 seconds in duration and comprised of repetition of up to 30 frames per second.

4.2. Performance Evaluation

To assess the classifier's performance in detail, various measures such as accuracy, precision, and recall were used to identify how the classifier performed, giving a sense of understanding. The evaluation, structured in two parts, showed that the CNN method achieves a high accuracy of 90.2% over the Shakefive2 dataset and 92.3% over the UoL dataset indicating that the proposed approach can be used in real-world applications.

The Shakefive2 dataset classification outcome, in terms of precision, recall, and F1-score, is demonstrated

as follows Table 2 and in Figure 10. In Table 2 Shakefive2, there were significant precisions/recall trade-offs. Actions like Fist-bump (precision=0.92, recall=0.86) and Pass-object (precision=1.00, recall=0.97) had high precision but low recall, and this shows that, though the model was accurate in predicting these interactions where it did, it still produced a large number of false negatives. In contrast, Shake-hand (precision=0.67, recall=0.83) and High-five (precision=0.82, recall=0.92) had a better recall, but at the cost of lower precision, since the two classes had a tendency to misclassify each other because of similar arm movements and postures.

Table 2. Detailed results of proposed system classification for the shakefive2 dataset.

Table classes	Precision	Recall	F1-score
Explain	0.80	0.90	0.82
Fist-bump	0.81	0.92	0.86
Shake-hand	0.67	0.83	0.84
High-five	0.82	0.92	0.80
Hug	0.79	0.83	0.83
Pass-object	1.00	0.97	0.97
Rock-paper-sizer	0.97	0.97	0.96
Thumbs-up	0.90	0.85	0.87

Figures 10 and 12 show the CNN recognition of human interaction results. Confusion matrix of the Shakefive2 dataset provides the systematic evaluation of the model ability to distinguish between multiple classes of human interactions. The matrix shows that the classification accuracy is high across the majority of the classes, such as explain, high five, fist bump, rock paper sizer and pass object, where the diagonal value is over 0.8 (and 1.00 in some cases). However, there are misclassifications, especially between Explain, Hug, and Thumbs-up, in which the model in some cases mixes up the classes. Explain is incorrectly classified as Thumbs-up with the rate of 0.10. Similarly, the High-Five and Fist-bump are likely to be incorrectly labeled as one another at the rate of 0.08, which is probably due to the similar arm movement pattern and postures. There are also misclassifications between Shake-hands and Fist-bump, which indicates that it is challenging to define these gestures, and both of them have a similar arm movement. These findings underscore the challenge of recognizing subtle differences in interactions involving similar movements or postures, which can lead to frequent misclassifications in these cases.

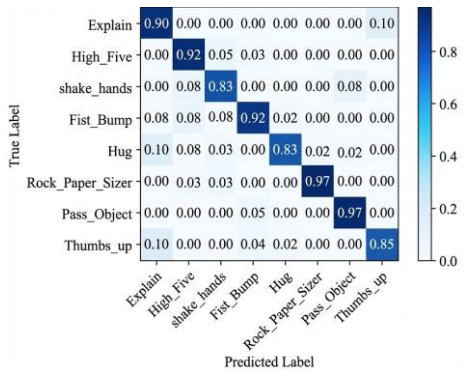


Figure 10. Confusion matrix result on shakefive2 dataset.

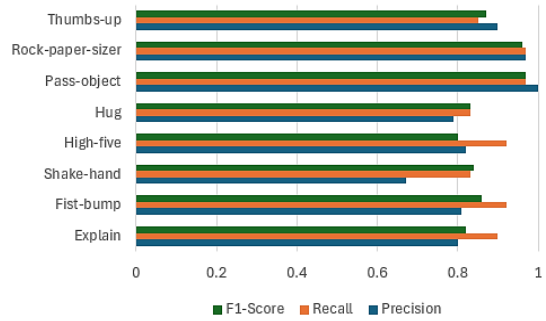


Figure 11. Recall, precision, and F1-score for each class on shakefive2 dataset.

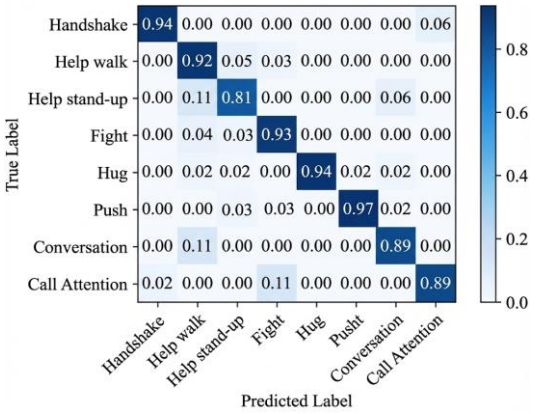


Figure 12. Human interaction recognition comparison with state-of-the-art methods on UoL dataset.

As the confusion matrix of the UoL dataset Figure 12 shows, the classification accuracy is high: 0.94 in Handshake, 0.92 in Help walk, 0.93 in Fight, 0.94 in Hug, 0.97 in Push. However, misclassifications occur between those behaviours exhibiting similar kinematic patterns. Help walk is also commonly confused with Help stand-up (0.05), because both have similar poses and leg actions. Similarly, Hug and Push have similar arm trajectories but body orientation is different resulting in a mislabeling rate of 0.06. Moreover, Conversation and Call Attention have an overlap; the former is incorrectly classified as the former (0.11) due to similar gestures of the upper body. These numbers show that the model has a strong performance, but it is unable to distinguish between those types of interactions where there is a slight difference in posture or dynamics of the movements.

Figure 13. Precision, recall, and F1-score for each class on UoL dataset.

Table 3. Detect the proposed system classification results for the UoL dataset.

Table classes	Precision	Recall	F1-score
Handshake	0.92	0.94	0.94
Help walk	0.87	0.92	0.89
Help stand-up	0.87	0.81	0.84
Fight	0.88	0.93	0.85
Hug	0.92	0.94	0.86
push	0.86	0.97	0.93
Conversation	0.89	0.89	0.99
Call attention	0.93	0.89	0.87

Figure 11 and 13 showing the comparison of each class with respect precision, recall, and F1-score for each class of dataset. In Table 3, the UoL dataset showed relatively good results, with high values of precision and recall of interactions with Handshake (precision=0.92, recall=0.94) and Hug (precision=0.92, recall=0.94). However, some of the classes that were considered similar in terms of visual appearances exhibited corrupted precision: Help walk (precision=0.87, recall=0.92) and Help stand-up (precision=0.87, recall=0.81) were frequently confused due to similar spatial and temporal characteristics. Push performed well with respect to the recall (precision=0.97, recall=0.97), but the precision was lower, and it implied that it misclassified the data to some extent. Similar problems were also faced by the Fight class (precision=0.88, recall=0.93) probably due to the similarity of aggressive gestures in interactions. The misclassifications were also mediated by differences in posture, lighting, and movement pattern specific to the subject, which restricted the model to perceive slight differences in gestures.

Table 4 show the results of an ablation study to evaluate the contribution of each component in our proposed human interaction recognition system. The ablation study results show the importance of each component in high performance, which represents the Tabular ablation study, all the components of the models have a substantial impact on the performance of the system. Most noticeably, leaving out CNN resulted in a drop in performance to 79% which proves that it plays a critical role in feature extraction and classification. When Texton maps were not used, a similar drop, namely, that of 92.3 percent to 82 percent was registered, which indicates the role played by Texton maps in the extraction of texture-based features. The removal of skeleton features, MOT and ViBe or QDA led to less significant decreases in performance, with accuracy being relatively constant.

Table 4. Detect the proposed system classification results for the shakefive2 dataset.

Experiments	HSV and median	MOT and ViBe	Texton map	Skeleton features	QDA	CNN	UoL	Shake five2
Full model	✓	✓	✓	✓	✓	✓	92.3%	90.2%
Without HSV and median filter	X	✓	✓	✓	✓	✓	87%	82%
Without MOT and ViBe	✓	X	✓	✓	✓	✓	84%	77%
Without Texton map	✓	✓	X	✓	✓	✓	82%	84%
Without skeleton features	✓	✓	✓	X	✓	✓	81%	86%
Without QDA	X	✓	✓	✓	X	✓	83%	79%
Without CNN	✓	✓	✓	✓	✓	X	79%	73%

Figures 14 and 15 give graphical representations of the evaluation outcomes of Shakefive2 and UoL, respectively. The bar chart represents the accuracy of each model of the experiments, and each bar represents an accuracy after the removal of a certain component. These bar charts clearly highlight the significant performance drop when CNN and Texton maps were removed, reinforcing their critical roles in the system.

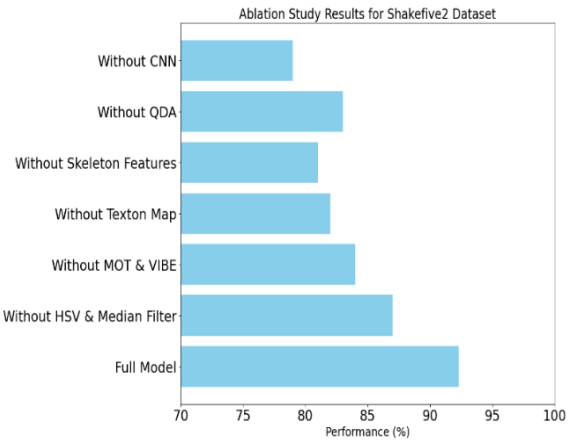


Figure 14. The performance of Shakefive2, with and without component ablations, and the comparison of the results.

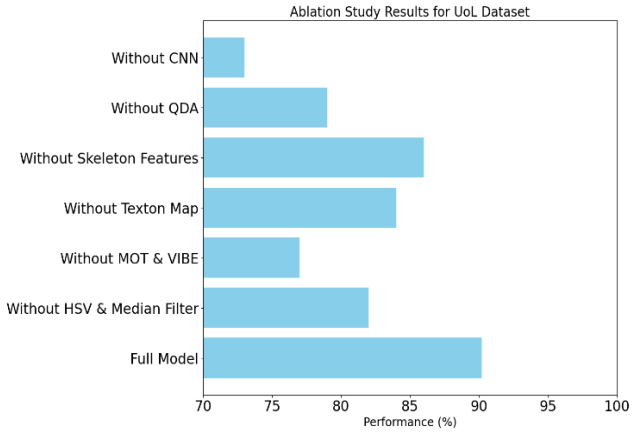


Figure 15. The performance of UoL with and without component ablations, and the comparison of the results.

Confidence intervals of accuracy, precision and recall were created using bootstrapping on UoL and shakefive2 datasets. The process is described in Table 5 and Figure 16 on the UoL data set and Table 6 and Figure 17 on the Shakefive2. Table 7 shows that our framework's performance was tested on shakefive2 and UoL datasets and the obtained accuracy was 90.2% and 92.3% respectively. Compared with the existing

interaction recognition methods, it can be deduced that our proposed framework had enhanced accuracy and performance. Our framework was more accurate in recognizing interactions across the video data than the previously proposed methods. Framework techniques were chosen specifically to cover some important aspects of appearance and motion to improve the proposed framework's abilities for interaction classification.

Table 5. The confidence intervals for accuracy, precision, and recall have been calculated using bootstrapping UoL dataset.

Metric	Mean value (%)	95% CI lower bound	95% CI upper bound
Accuracy	92.3	90	93
Precision	91	89.5	92.5
Recall	90	88	91.5

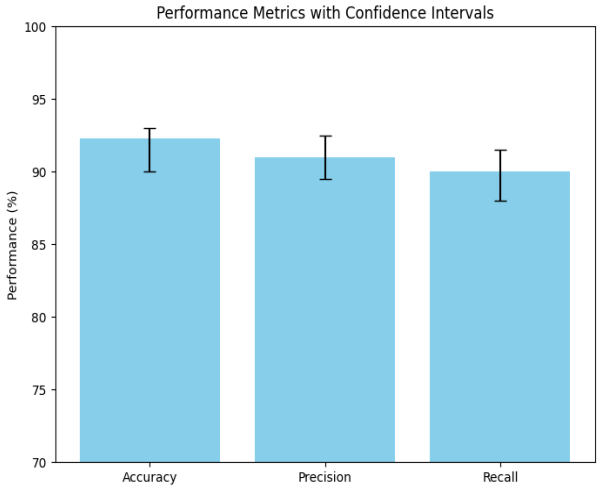


Figure 16. The confidence intervals for accuracy, precision, and recall have been calculated using bootstrapping of UoL dataset.

Table 6. The confidence intervals for accuracy, precision, and recall have been calculated using bootstrapping of UoL dataset.

Metric	Mean value (%)	95% CI lower bound	95% CI upper bound
Accuracy	90.20	89.20	92.30
Precision	89.25	77.5	91.5
Recall	89.87	86.37	94.0

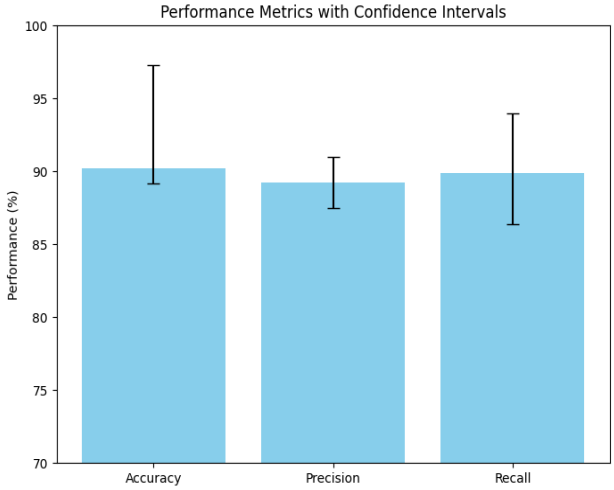


Figure 17. The confidence intervals for accuracy, precision, and recall have been calculated using bootstrapping of shakefive2 dataset.

Table 7. Comparison with other SR methods over RGB-d dataset.

Dataset	Methods	Accuracy (%)
Shakefive2 dataset	Deformable parts models [34]	65.0%- 87.0%
	Histogram of Oriented Gradients (HOG) and Motion Boundary Histogram (MBH) [35]	82.0%
	Proposed	**90.2%**
UoL dataset	Statistical and geometrical features [36]	85.5%
	Probabilistic merging of fusion based features [38]	86.2%
	SVM [39]	87.0%
	Proposed	**92.3**

5. Conclusions

This paper has developed an effective CNN-based framework for an interaction recognition approach to recognize and categorize interactions. The proposed method includes using HSV color transformation, object silhouette extraction using MOT or ViBe methods, the Texton maps and skeleton features as the features extraction and the QDA as the feature separation. The experimentation on the Shakefive2 and UoL datasets shows that the suggested approach is effective. The method got an 84% recognition rate on the Shakefive2 dataset and 87% accuracy on the UoL dataset, which are the best results to date in interaction recognition. The method has been proven to be accurate in distinguishing interactions based on the distinctive features of each interaction. The results show the effectiveness and power of the method in practice based on CNN and implemented into the system. The proposed technique which successfully combines various methods is a realistic approach that solves computer vision and human interaction recognition real-world problems. The following research is the study of more advanced techniques that can be used to optimize this method and improve accuracy by exploring many other datasets. The proposed method can also be extended to address complex interaction recognition tasks and could be applied to various domains, where accurate interaction identification is a must. Conclusively, the CNN-based interaction recognition approach discussed in this paper offers a solid basis to interaction recognition development and makes a relevant contribution to the creation of intelligent systems that could understand and interpret the activities of people in the current applications.

Funding

This work was supported through Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] Al Mokhtar Z. and Dawwd S., "3D VAE Video Prediction Model with Kullback Leibler Loss Enhancement," *The International Arab Journal of Information Technology*, vol. 21, no. 5, pp. 879-

888, 2024. DOI:10.34028/iajit/21/5/9

[2] Alonazi M., Ansar H., Al Mudawi N., Alotaibi S., and et al., "Smart Healthcare Hand Gesture Recognition Using CNN-Based Detector and Deep Belief Network," *IEEE Access*, vol. 11, pp. 84922-84933, 2023. DOI: 10.1109/ACCESS.2023.3289389

[3] Barnich O. and Droogenbroeck M., "ViBe: A Universal Background Subtraction Algorithm for Video Sequences," *IEEE Transactions on Image Processing*, vol. 20, pp. 1709-1724, 2010. https://doi.org/10.1109/TIP.2010.2101613

[4] Chaaraoui A., Perez P., and Revuelta F., "A Review on Vision Techniques Applied to Human Behaviour Analysis for Ambient-Assisted Living," *Expert Systems with Applications*, vol. 39, pp. 10873-10888, 2012. https://doi.org/10.1016/j.eswa.2012.03.005

[5] Cheng X. and Zhang P., "Enhanced Soccer Training Simulation Using Progressive Wasserstein GAN and Termite Life Cycle Optimization in Virtual Reality," *The International Arab Journal of Information Technology*, vol. 21, no. 4, pp. 549-559, 2024. DOI: 10.34028/iajit/21/4/1

[6] Coppola C., Cosar S., Faria D., and Bellotto N., "Automatic Detection of Human Interactions from RGB-D Data for Social Activity Classification," *in Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication*, Lisbon, pp. 871-876, 2017. https://doi.org/10.1109/ROMAN.2017.8172405

[7] Dua N., Singh S., and Semwal V., "Multi-Input CNN-GRU Based Human Activity Recognition Using Wearable Sensors," *Computing*, vol. 103, pp. 1461-1478, 2021. https://doi.org/10.1007/s00607-021-00928-8

[8] Feudo S., Dion J., Renaud F., Kerschen G and Noel J., "Video Analysis of Nonlinear Systems with Extended Kalman Filtering for Modal Identification," *Nonlinear Dynamics*, vol. 111 pp. 13263-13277, 2023. https://link.springer.com/article/10.1007/s11071-023-08560-1

[9] Garcia S., Baena C., and Salcedo A., "Human Activities Recognition Using Semi-Supervised SVM and Hidden Markov Models," *TecnoLogicas*, vol. 26, no. 56, pp. 1-19, 2023. https://doi.org/10.22430/22565337.2474

[10] Gemeren C., Poppe R., and Veltkamp R., "Hands-on: Deformable Pose and Motion Models for Spatiotemporal Localization of Fine-Grained Dyadic Interactions," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 16, pp. 1-16, 2018. https://jivp-eurasipjournals.springeropen.com/articles/10.1186/s13640-018-0255-0

[11] Gemeren C., Poppe R., and Veltkamp R., DPM Configurations for Human Interaction Detection, Utrecht University, 2016. https://webspace.science.uu.nl/~veltk101/publications/art/nccv2015_p35L.pdf

[12] Hasan R. and Alani N., "A Comparative Analysis Using Silhouette Extraction Methods for Dynamic Objects in Monocular Vision," *Cloud Computing and Data Science*, vol. 1, pp. 1-12, 2022. https://www.researchgate.net/publication/357860149_A_Comparative_Analysis_Using_Silhouette_Extraction_Methods_for_Dynamic_Objects_in_Monocular_Vision

[13] He L., Jiang D., Yang L., Pei E., and et al., "Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks," *in Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, Brisbane, pp. 73-80, 2015. https://doi.org/10.1145/2808196.2811641

[14] Hua G., Kumar H., Aradhya M., and Maheshan M., "A New Effective Speed and Distance Feature Descriptor Based on Optical Flow Approach in HAR," *Revue D'Intelligence Artificielle*, vol. 37, pp. 109-115, 2023. http://dx.doi.org/10.18280/ria.370114

[15] Jalal A., Kim Y., Kim Y., Kamal S., and et al., "Robust Human Activity Recognition from Depth Video Using Spatiotemporal Multi-Fused Features," *Pattern Recognition*, vol. 61, pp. 295-308, 2017. https://doi.org/10.1016/j.patcog.2016.08.003

[16] Khan A., Chefranov A., and Demirel H., "Image Scene Geometry Recognition Using Low-Level Features Fusion at Multi-Layer Deep CNN," *Neurocomputing*, vol. 440, pp. 111-126, 2021 https://doi.org/10.1016/j.neucom.2021.01.085

[17] Khodabandelou G., Moon H., Amirat Y., and Mohammed S, "A Fuzzy Convolutional Attention-Based GRU Network for Human Activity Recognition," *Engineering Applications of Artificial Intelligence*, vol. 118, pp. 105702, 2022. https://doi.org/10.1016/j.engappai.2022.105702

[18] Koping L., Shirahama K., and Grzegorzek M., "A General Framework for Sensor-Based Human Activity Recognition," *Computers in Biology and Medicine*, vol. 95, pp. 248-260, 2018. https://doi.org/10.1016/j.compbiomed.2017.12.025

[19] Liu X., Shi H., Hong X., Chen H., and et al., "Hidden States Exploration for 3D Skeleton-Based Gesture Recognition," *in Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, pp. 1846-1855, 2019. https://doi.org/10.1109/WACV.2019.00201

[20] Luo W., Xing J., Milan A., Zhang S., and et al., "Multiple Object Tracking: A Literature Review," *Artificial Intelligence*, vol. 293, pp. 103448, 2021. https://doi.org/10.1016/j.artint.2020.103448

[21] Luvizon D., Tabia H., and Picard D., "Learning Features Combination for Human Action Recognition from Skeleton Sequences," *Pattern Recognition Letters*, vol. 99, pp. 13-20, 2017. https://doi.org/10.1016/j.patrec.2017.02.001

[22] Manzi A., Fiorini L., Limosani R., Dario P., and Cavallo F., "Two-Person Activity Recognition Using Skeleton Data," *IET Computer Vision*, vol. 12, no. 1, pp. 27-35, 2018. https://doi.org/10.1049/iet-cvi.2017.0118

[23] Modi N. and Ramakrishna M., "An Investigation of Camera Movements and Capture Techniques on Optical Flow for Real-Time Rendering and Presentation," *Journal of Real-Time Image Processing*, vol. 20, no. 60, pp. 1-15, 2023. https://doi.org/10.1007/s11554-023-01322-7

[24] Mukherjee S., Anvitha L., and Lahari M., "Human Activity Recognition in RGB-D Videos by Dynamic Images," *Multimedia Tools and Applications*, vol. 79, pp. 19787-19801, 2020. https://doi.org/10.48550/arXiv.1807.02947

[25] Nadeem A., Jalal A., and Kim K., "Accurate Physical Activity Recognition Using Multidimensional Features and Markov Model for Smart Health Fitness" *Symmetry*, vol. 12, no. 11, pp. 1-17, 2020. https://doi.org/10.3390/sym12111766

[26] Pang Y., Ke Q., Rahmani H., Bailey J., Liuand J., "IGFormer: Interaction Graph Transformer for Skeleton-Based Human Interaction Recognition," *in Proceedings of the European Conference on Computer Vision*, Tel Aviv, pp. 605-622. https://arxiv.org/abs/2207.12100

[27] Pareek P. and Thakkar A., "A Survey on Video-Based Human Action Recognition: Recent Updates, Datasets, Challenges, and Applications," *Artificial Intelligence Review*, vol. 54, pp. 2259-2322, 2021. https://doi.org/10.1007/s10462-020-09904-8

[28] Saleem G., Bajwa U., and Raza R., "Toward Human Activity Recognition: A Survey," *Neural Computing and Applications*, vol. 35, pp. 4145-4182, 2023. https://doi.org/10.1007/s00521-022-07937-4

[29] Samir H., Abd El Munim H., and Aly G., "Suspicious Human Activity Recognition Using Statistical Features," *in Proceedings of the 13th International Conference on Computer Engineering and Systems*, Cairo, pp. 589-594, 2018. https://doi.org/10.1109/ICCES.2018.8639457

[30] Shi L., Zhang Y., Cheng J., and Lu H., "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, pp. 12026-12035, 2019. https://doi.org/10.1109/CVPR.2019.01230

[31] Shuvo M., Ahmed N., Nouduri K., and Palaniappan K., "A Hybrid Approach for Human Activity Recognition with Support Vector Machine and 1D Convolutional Neural Network" *in Proceedings of the IEEE Applied Imagery Pattern Recognition, Workshop*, Washington (DC), pp. 1-5, 2020. https://doi.org/10.1109/AIPR50011.2020.9425332

[32] Waheed M., Javeed M., and Jalal A., "A Novel Deep Learning Model for Understanding Two-Person Interactions Using Depth Sensors," *in Proceedings of the International Conference on Innovative Computing*, Lahore, pp. 1-8, 2021. https://doi.org/10.1109/ICIC53490.2021.9692946

[33] Wang X., Sun Z., Chehri A., Jeon G., and *et al*., "Deep Learning and Multi-Modal Fusion for Real-Time Multi-Object Tracking: Algorithms, Challenges, Datasets, and Comparative Study," *Information Fusion*, vol. 105, pp. 102247, 2024. https://doi.org/10.1016/j.inffus.2024.102247

[34] Yadav S., Tiwari K., Pandey H., and Akbar S., "A Review of Multimodal Human Activity Recognition with Special Emphasis on Classification, Applications, Challenges and Future Directions," *Knowledge-Based Systems*, vol. 223, pp. 106970, 2021. https://doi.org/10.1016/j.knosys.2021.106970

[35] Yan S., Xiong Y., and Lin D., "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," *in Proceedings of the AAAI Conference on Artificial Intelligence*, Lousiana, pp. 7444-7452, 2018. https://doi.org/10.1609/aaai.v32i1.12328

[36] Yi M., Lee W., and Hwang S., "A Human Activity Recognition Method Based on Lightweight Feature Extraction Combined with Pruned and Quantized CNN for Wearable Device," *IEEE Transactions on Consumer Electronics*, vol. 69, pp. 657-670, 2023. https://doi.org/10.1109/TCE.2023.3266506

[37] Zhang L., "Enterprise Employee Work Behavior Recognition Method Based on Faster Region Convolutional Neural Network," *The International Arab Journal of Information Technology*, vol. 22, no. 2, pp. 291-302, 2025. https://doi.org/10.34028/iajit/22/2/7

[38] Zhang S., Li Y., Zhang S., Shahabi F., and et al., "Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances," *Sensors*, vol. arXiv:2111.00418v5, pp. 1-42, 2020. https://arxiv.org/abs/2111.00418v5

[39] Zheng L., Tang M., Chen Y., Zhu G., and et al., "Improving Multiple Object Tracking with Single Object Tracking," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, pp. 2453-2462, 2021. https://doi.org/10.1109/CVPR46437.2021.00248

Tanvir Fatima Naik Bukht is a Lecturer and dedicated and passionate researcher pursuing a Ph.D. at Air University in Islamabad, Pakistan. Simultaneously, she works as a lecturer in the CGD Department at the same university. With an impressive academic background, she holds an MPhil degree from the Institute of Southern Punjab in Multan, Pakistan, and a Master's in Computer Science from the Virtual University of Pakistan. Tanvir's research has primarily focused on the cutting-edge fields of Cyber Security, Information Security, Internet of Things Security, and Computer Vision. Additionally, she has delved into Digital Image Processing and Medical Imaging Challenges in Information Systems, contributing valuable insights to the scientific community. Throughout her academic journey, Tanvir has displayed a keen interest in staying updated with Technological Advancements and has collaborated with various researchers in her field. As a result, she has published several high-quality research papers in reputable international journals, including those indexed by IEEE Access.

Haita Alhaston is Associate Professor at Umm Al-Qura University, passionate about Bioimage Informatics Digital Transformation and Innovation working as an AI and Data Scientist, Med AI Mentor, AI Consultant and Trainer IEEE member.

Noif Alshamrari is an Assistant Professor within the College of Computer and Information Sciences, Majmaah University (KSA). He is currently the Vice Dean of the Deanship of Information Technology and E-Learning, Majmaah Universtiy. His research interests include Object Detection and Segmentation, within Automotive Applications, using state of the art technique Convolutional Neural Networks (CNNs). Dr. Alshammari holds a PhD in computer vision from Durham University (UK).

Nouf Abdullah Almujally received the Ph.D. in Computer Science from the University of Warwick, UK. She is currently an Assistant Professor in Computer Science at the Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University (PNU), Riyadh, Saudi Arabia. Her research interests include Human-Computer Interaction (HCI), Artificial Intelligence (AI), Machine Learning, Deep Learning, and Computer Based Applications.

Ahmad Jalal is currently an Associate Professor from Department of Computer Science and Engineering, Air University, Pakistan. He received his Ph.D. degree in the Department of Biomedical Engineering at Kyung Hee University, Republic of Korea. Now, he was working as Post-Doctoral Research fellowship at POSTECH. His research interest includes Multimedia Contents, Artificial Intelligence.