Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.041335

ARTICLE

A Novel Human Interaction Framework Using Quadratic Discriminant Analysis with HMM

Tanvir Fatima Naik Bukht[1], Naif Al Mudawi[2], Saud S. Alotaibi[3], Abdulwahab Alazeb[2], Mohammed Alonazi[4], Aisha Ahmed AlArfaj[5], Ahmad Jalal[1] and Jaekwang Kim[6,*]

[1]Department of Computer Science, Air University, Islamabad, 44000, Pakistan

[2]Department of Computer Science, College of Computer Science and Information System, Najran University, Najran, 55461, Saudi Arabia

[3]Information Systems Department, Umm Al-Qura University, Makkah, Saudi Arabia

[4]Department of Information Systems, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, 16273, Saudi Arabia

[5]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia

[6]Convergence Program for Social Innovation, Sungkyunkwan University, Suwon, 03063, Korea

*Corresponding Author: Jaekwang Kim. Email: linux@skku.edu

Received: 18 April 2023 Accepted: 11 August 2023 Published: 29 November 2023

ABSTRACT

Human-human interaction recognition is crucial in computer vision fields like surveillance, human-computer interaction, and social robotics. It enhances systems' ability to interpret and respond to human behavior precisely. This research focuses on recognizing human interaction behaviors using a static image, which is challenging due to the complexity of diverse actions. The overall purpose of this study is to develop a robust and accurate system for human interaction recognition. This research presents a novel image-based human interaction recognition method using a Hidden Markov Model (HMM). The technique employs hue, saturation, and intensity (HSI) color transformation to enhance colors in video frames, making them more vibrant and visually appealing, especially in low-contrast or washed-out scenes. Gaussian filters reduce noise and smooth imperfections followed by silhouette extraction using a statistical method. Feature extraction uses the features from Accelerated Segment Test (FAST), Oriented FAST, and Rotated BRIEF (ORB) techniques. The application of Quadratic Discriminant Analysis (QDA) for feature fusion and discrimination enables high-dimensional data to be effectively analyzed, thus further enhancing the classification process. It ensures that the final features loaded into the HMM classifier accurately represent the relevant human activities. The impressive accuracy rates of 93% and 94.6% achieved in the BIT-Interaction and UT-Interaction datasets respectively, highlight the success and reliability of the proposed technique. The proposed approach addresses challenges in various domains by focusing on frame improvement, silhouette and feature extraction, feature fusion, and HMM classification. This enhances data quality, accuracy, adaptability, reliability, and reduction of errors.

KEYWORDS

Human interaction recognition; HMM classification; quadratic discriminant analysis; dimensionality reduction

1 Introduction

Human interaction recognition (HIR) in computer vision refers to a system's ability to recognize and recognize the gestures, expressions, and movements of humans engaged in face-to-face communication. This technology enables computers to comprehend and respond to human behaviour, allowing for more natural and intuitive interactions between humans and machines. Human-to-human interaction in computer vision applications includes video conferencing, virtual reality, and gaming. This technology could revolutionize how we connect with technology and one another by allowing computers to understand human behaviour.

HIR is a challenging computer vision problem that seeks to comprehend human behavior by analyzing visual data such as photos and videos. The goal is to recognize complex, human-to-human interactions; however, this is difficult due to challenges such as viewpoint fluctuation, occlusion, ambiguity, data scarcity, and interaction complexity. As a result, the performance and application of most existing HIR approaches are limited. HIR advancements could allow applications such as better video/image surveillance, improved human-computer interactions, and safer intelligent modes of transport. Image-based interaction recognition is more challenging than video-based action detection due to limited data, the blurry background of the images, ambiguous qualities like Similar looking interactions may have different meanings, and training data like hard-to-code large amounts of marked-up data of human interactions and Complex interactions. This has led to an interest in human interaction recognition, to its diverse applications, such as human-machine interaction, behavioral biometrics, surveillance systems, environmental intelligence, assisted living, and human-computer interaction. However, this research aims to develop a novel image-based human interaction recognition method that can improve performance compared to existing methods. This research is particularly significant as it has the potential to contribute to various computer vision applications such as video surveillance, human-computer interaction, and intelligent transportation systems.

Several methods have been proposed for recognizing human interactions, including histogram of oriented gradients (HOG), local binary patterns (LBP), and deep neural networks DNNs [1–6]. However, these methods have limitations and may not always provide optimal results. The following are some limitations of HOG and LBP techniques, which rely on handcrafted features that may not be relevant to a wide range of human interaction contexts. The features of each new dataset must be carefully designed and optimized. DNNs require a huge amount of labeled training data for human interaction recognition, which can be costly and time-consuming to acquire. DNNs have considerable computing requirements during training and inference, which limits their use in resource-constrained situations. HOG/LBP and standard DNNs do not explicitly recreate the temporal dynamics of human interactions, which are essential for successful recognition.

To address these limitations, we present a model that represents the developmental process of human interactions using Hidden Markov Models. Compared to other methods, HMMs require fewer parameters to train and can obtain good results with smaller data sets for training. Because HMM has modest computing needs, they are appropriate for real-world applications. It can demonstrate temporal interconnectedness and the development of human interactions through time. This study offers a unique human interaction recognition method that combines the best features of QDA and the HMM. QDA is an effective feature fusion and discriminating tool for high-dimensional data analysis. HMM is a statistical model successfully applied to human interaction recognition. It can be used to classify sequential data. Pre-processing the photos with HSI color transformation to boost contrast and Gaussian filters to reduce noise is the first step in our suggested approach for human interaction recognition. Following that, statistical approaches are utilized to extract the person's silhouette. The

FAST and ORB approaches are used to extract features. These features are then passed through QDA for feature fusion and discrimination. Finally, HMM is used to classify the features into appropriate human activities. The proposed method achieved an accuracy of 93% on a dataset of human activities and shows potential for improving human recognition performance. The HIR system can recognize complex human activities, such as shaking hands, hugging, kicking, patting, pushing, hifi, bending, and boxing. The main contributions of our proposed model:

- Frame enhancement and extraction using HSI transformation and Gaussian filter.
- Silhouette extraction using statistical methods.
- Feature extraction using FAST and ORB.
- Quadratic Discriminant Analysis technique is applied for feature fusion and discrimination, which are later identified using the Hidden Markov model.

The suggested framework's compact design makes it suitable for deployment on any edge device. There is no need for waiting or extra processing time; everything can be done in real-time. Complexity and computational requirements may make the suggested approach suitable for large datasets. Noise or missing values in input data may influence the algorithm's accuracy and dependability. The algorithm may require significant domain-specific knowledge and experience to implement and configure.

The research paper is organized into the following sections: Section 2 provides a detailed review of the related work in the field of HIR. Section 3 focuses on the design and structure of the proposed system, explaining the methodology and techniques used for image dataset pre-processing, feature extraction, and classification. Section 4 presents the experimental analysis and results of the proposed method, including the system's accuracy and efficiency in recognizing different human activities. Finally, Section 5 concludes the research paper by summarising.

2 Related Work

Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have recently emerged as promising candidates for human recognition thanks to advancements in deep learning [7]. These methods can learn hierarchical representations of images and videos that can be used to recognize activities. CNN-based methods have shown impressive performance on various datasets and benchmarks but require a large amount of labeled data for training and have a high computational cost. Other handcrafted feature extraction techniques like Motion History Image (MHI) [8], Optical Flow, and 3D convolutional neural networks (3DCNNs) have also been proposed for HIR. Furthermore, certain hybrid approaches have been presented in the literature to enhance the performance of human action identification, such as integrating CNN with HMM [9]. In this setup, the CNN is utilized to extract the features from the image, while the HMM models the temporal information of the activities.

2.1 Image-Based HIR

Human interaction recognition recognizes and classifies human activities using visual data, such as images or videos, to recognize and organize them. Regarding recognizing human-human interactions (HHI) in computer vision, researchers have proposed various methods that use machine learning classifiers such as Random Forest, Support Vector Machine (SVM), Decision Trees, and HMM to recognize and categorize human interactions, such as shake_hands, hugs, and high-fives, into distinctly different classes [7,8]. These methods extract handcrafted elements, including spatial-temporal information, posture, and gesture, from images or videos and use classifiers to recognize and classify interactions. Human-object interaction (HOI) recognition analyses human-object relationships in images and videos. They recognize and classify activities like carrying a bag or sitting on a chair. Researchers employ machine learning classifiers like Random Forest, SVM, Decision Trees, and HMM to recognize and classify these interactions by extracting information like object location, size, and object-human interaction from images and videos.

Histograms of oriented gradients are one of the most popular approaches to human action recognition in images [9–11], which can extract features from stable images against modifications like brightness and contrast. Following feature collection, a classifier like an SVM or a linear discriminant analysis (LDA) is trained to recognize the activities based on the data collected LDA. HOG-based methods effectively recognize simple and repetitive activities but may not always provide optimal performance for more complex and varied activities. Another popular method for human interaction recognition is based on LBP [12,13], which encodes the local texture of the image. LBP-based methods have also been effective for recognizing simple and repetitive activities but effective for recognizing and recognizing simple and repetitive activities. However, they may not be as robust to changes in lighting and scale as HOG-based methods.

2.2 Markers-Based HIR

Markers-based Human Interaction Recognition (HIR) is a computer vision technique that uses physical markers, typically placed on specific body joints or landmarks, to identify and track human movements and interactions in real-time [14]. Markers accurately capture motion data, which algorithms may evaluate to recognize human activities and interactions. Virtual reality, gaming, sports, and rehabilitation use marker-based HIR for accurate motion analysis and intuitive user interfaces [15]. Although marker-based HIR produces high-quality data, it can be limited by the need for hardware, which may hinder the natural movements of the tracked subjects [16].

Numerous studies have examined marker-based HIR methods. Hidden Markov Models are famous for modeling and recognizing human behaviour [17]. HMMs are statistical models that can represent complex temporal patterns and have been effectively used in human activity recognition and motion analysis [18]. HMMs were employed to recognize tennis actions from marker-based motion capture data [19,20]. Deep learning techniques, such as CNNs and RNNs, have been investigated by other researchers for marker-based HIR [21]. These methods recognize complicated activities and interconnections better. They need substantial training datasets and computer power. HMMs, deep learning algorithms, and others have been studied to improve marker-based HIR recognition accuracy and efficiency.

3 Structure of Designed System

This research paper represents a novel approach to HIR. The suggested system consists of five steps. Stretching the frame contrast occurs during pre-processing. The subsequent process involves extracting the person's silhouette. In the third phase, two features are extracted using FAST and ORB and combined. The fourth step explores high-dimensional data features using the Quadratic Discriminant Analysis technique for feature fusion and discrimination. This feature is subsequently used to train the HMM for the last stage of interaction recognition. In Fig. 1, the suggested architecture is displayed.

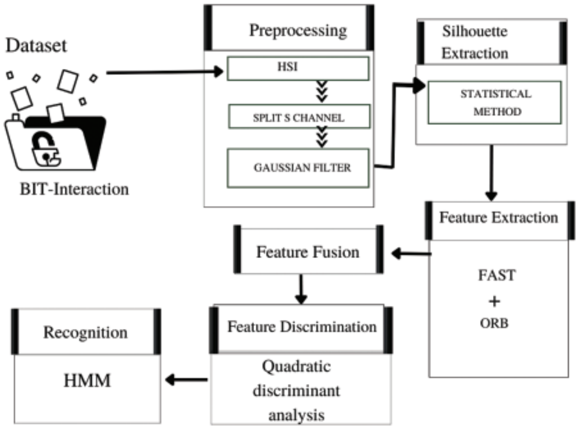


Figure 1: Flow diagram of the proposed HIR approach

3.1 Pre-Processing

Incorrect estimates of human behavior can be avoided with the help of pre-processing, which is why removing noise from the input frames is a crucial step in extracting essential features. In this study, we employ a simplified pre-processing solution consisting of two steps: (a) HSI color transformation and (b) picking the optimal channel and applying a Gaussian filter to target this problem.

3.1.1 Selecting the Optimal Channel

We apply the HIS transformation to the input video frame. The HSI effect deconstructs the source video frame, which we may divide into three channels (hue, saturation, and intensity) using the coordinates β(x,y). The red, green, and blue channels are represented by $\beta_r(x,y)$, $\beta_g(x,y)$, and $\beta_b(x,y)$, respectively, are normalized by dividing them by the sum of the three channels. The final HSI transformation is then calculated using the following equations:

$$Y_1 = \frac{1}{2}\left\{\left(\beta_r(x,y) - \beta_g(x,y)\right) + \beta_r(x,y) - \beta_b(x,y)\right\} \tag{1}$$

$$g_1 = \left(\beta_r(x,y) - \beta_g(x,y)^2\right), g_2 = \beta_r(x,y) - \beta_b(x,y) \times \left(\beta_g(x,y) - \beta_b(x,y)\right) \tag{2}$$

$$Y_2 = \sqrt{g_1 + g_2} \tag{3}$$

$$\phi = \cos^{-1}\left(\frac{Y_1}{Y_2 + eps}\right) \tag{4}$$

The hue channel is then calculated as follows:

$$\beta_H(x,y) = \cos^{-1}\left(\frac{Y_1}{\sqrt{Y_2}}\right) \tag{5}$$

Fig. 2 represents the HSI of the input image. The hue, saturation, and intensity channels are displayed as grayscale images, showing the original image's color information, vividness, and brightness. The figure highlights the importance of considering different color channels for image processing and analysis. It highlights each channel's key features and emphasizes the HSI representation's significance in image analysis.

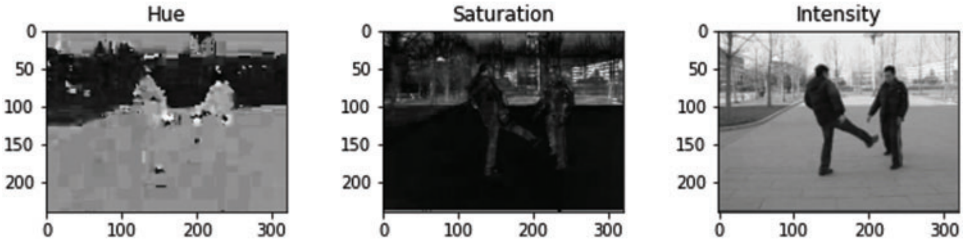


Figure 2: Hue-Saturation-Intensity (HSI) representation of the original image

3.1.2 Gaussian Filter

The Gaussian filter is a widely used image-processing technique in computer vision applications for smoothing or blurring images [19–24]. It is based on the Gaussian function, a bell-shaped curve that weights pixels in an image. To pre-process our image data in our study, we used the Gaussian filter in conjunction with the HSI transformation. We applied the Gaussian filter to each channel separately after performing the HSI transformation, which divides the original image into hue, saturation, and intensity channels. This lets us remove noise and high-frequency material from photos while retaining edges and other critical elements. Furthermore, because of the low-pass character of the filter, the Gaussian filter can be used to lessen the amount of aliasing in an image.

The Gaussian filter can be represented by Eq. (6):

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \tag{6}$$

where $G(x, y)$ is the Gaussian filter, e is Euler's number, and σ is the standard deviation of the Gaussian function. The standard deviation determines the spread of the Gaussian function and directly controls the amount of smoothing or blurring applied to the image. Gaussian filtering results are shown in Fig. 3.

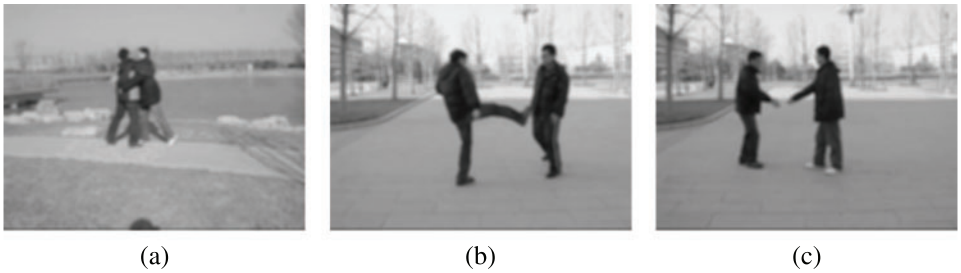


Figure 3: Gaussian filtering applied, which highlights the improved contrast and reduced noise, (a) show hug, (b) show kick, and (c) shake_hand

We have used this filter because it is a simple and effective technique for reducing image noise and preserving important features. It is crucial for accurately recognizing human activities in our research. By applying it separately to each channel after HSI, we can improve the performance of our image-based human interaction recognition method.

3.2 *Silhouette Extraction*

Silhouette extraction is crucial to object recognition, tracking, and segmentation in computer vision. Statistic-based silhouette extraction is durable and accurate. The Gaussian-Mixture-Model, Expectation-Maximization Algorithm, K-Means Clustering, Mean-Shift, and Spectral Clustering are used for image segmentation, object recognition, and data analysis [25–28]. These models use statistical data to get insights and perform various tasks.

In this work, we suggest using the Gaussian Mixture Model (GMM) to extract silhouettes. Our method performs thresholding and inverse thresholding on the image to produce a binary mask. After applying the mask, the GMM pulls out the silhouette from the image. The resulting silhouette is a two-dimensional binary image, with the foreground pixels denoting the subject and the background pixels indicating the background. The output is represented by a silhouette superimposed on top of the original color image on a black backdrop. The results are displayed in Fig. 4, demonstrating the accuracy and efficiency with which our method achieves silhouette extraction.

$$p\left(x\right) = \sum\nolimits_{i=1}^{k} w_i \phi\left(x; \mu_i, \Sigma_i\right) \tag{7}$$

Eq. (7), in GMM where $p\left(x\right)$ is the probability density function, w_i is the weight of the *ith* Gaussian component, ϕ is the Gaussian distribution function, k is the number of Gaussian components, μ_i is the mean vector, and Σ_i is the covariance matrix.

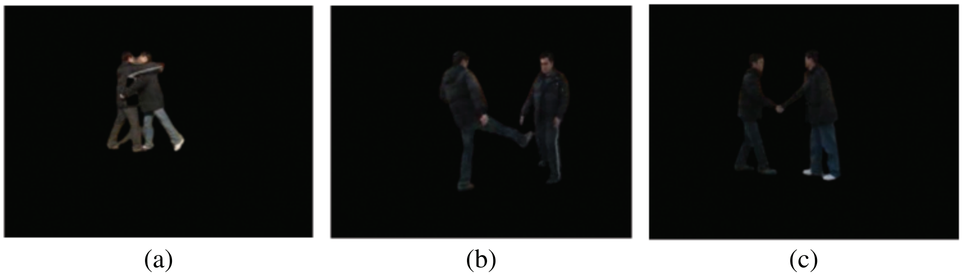

(a) (b) (c)

Figure 4: Silhouette extraction using a statistical method, (a) show hug, (b) show kick, and (c) shake_hand

The algorithm extracts an object's silhouette from an input image. It converts the image to grayscale, applies a GMM background subtractor, and thresholds the foreground mask. The mask becomes inverted to produce a binary picture if the mean value exceeds a certain threshold. The silhouette is shown on a black backdrop, with the original image overlaid with the silhouette.

Algorithm 1: Silhouette extraction using statistical method (GMM)

```
Input:   Images
Output:  A silhouette image on a black background and the original image with a silhouette overlay;
begin
    I = read_input_image()
    I_gray = convert_to_grayscale(I)
    GMM = initialize_gmm_background_subtractor()
    while True:
      M_t = GMM.apply(I_gray)
```

(Continued)

Algorithm 1 (continued)

```
μ_t = mean(M_t), Compute the mean value of the foreground mask
if μ_t > T:
    B_t = threshold(M_t, T, 255, THRESH_BINARY)
    B_t_inv = bitwise_not(B_t)
    S_t = bitwise_and(I, I, mask=B_t_inv)
      show_image_on_black_background(S_t)
      show_original_image_with_silhouette(I, S_t)
  else:
      Continue
  end if
Return
```

3.3 Feature Extraction

Despite its ubiquity and effectiveness, the FAST algorithm still has certain shortcomings. One of the key limitations of the FAST algorithm is its sensitivity to noise and complexity in images. The combined usage of the FAST and ORB extractors for features is well recognized as an excellent method for imagining feature extraction. The FAST feature detector has been recognized for its great computational speed and repeatability, but the ORB descriptor is known for its resistance to scale and rotation changes. Combining these two strategies yields a robust feature extraction strategy that can handle many tough conditions. Using FAST and ORB, the approach can encapsulate high-speed computing and resistance to diverse transformations. Furthermore, utilizing ORB descriptors can assist in reducing the number of false matches in the matching stage, enhancing accuracy and efficiency.

3.3.1 Features from Accelerated Segment Test (FAST)

The FAST Algorithm is a popular technique for detecting and extracting key characteristics from digital images. Many computer vision applications, such as object recognition, tracking, and registration, rely on feature detection. Extracting features from images accurately and effectively is critical in many computer vision systems. Feature detection algorithms generally seek out identifiable and recurrent structures in photos. These keypoint structures define distinct points in the image that can be used to identify and track things across multiple frames or images.

FAST algorithm works by comparing a pixel intensity value with the values of its surrounding pixels in a circular pattern. If a certain number of contiguous pixels have intensities that are either higher or lower than the central pixel, then the central pixel is marked as a corner. First, the algorithm calculates the difference between the intensities of the pixel at the center and its surrounding pixels, with a radius of 3 pixels:

$$d_i = I_{\{p_i\}} - I_{\{pc\}} \tag{8}$$

where p_c is the center pixel and p_i are the surrounding pixels.

Next, the algorithm selects a threshold value T, and a pixel p_c is considered a corner if there exist n contiguous pixels in the circle around p_c whose intensities are all greater than $I_{\{pc\}} + TorlessthanI_{\{pc\}} - T$:

$$\mathrm{C} = p_c, |, \exists, \mathrm{n}, p_c \text{ with } I_{\{p_i\}} > I_{\{pc\}} + \mathrm{T} \text{ or } I_{\{p_i\}} < I_{\{pc\}} - \mathrm{T} \tag{9}$$

Finally, to speed up the algorithm, the threshold value T is calculated as a fraction k of the maximum intensity range:

$$T = k \cdot max_{p \in I} I_P - min_{p \in I} I_P \tag{10}$$

where I is the image and k is a defined constant.

The results presented in Fig. 5 were obtained after extracting features from the images using FAST. The detected features were then used for subsequent analysis and evaluation.

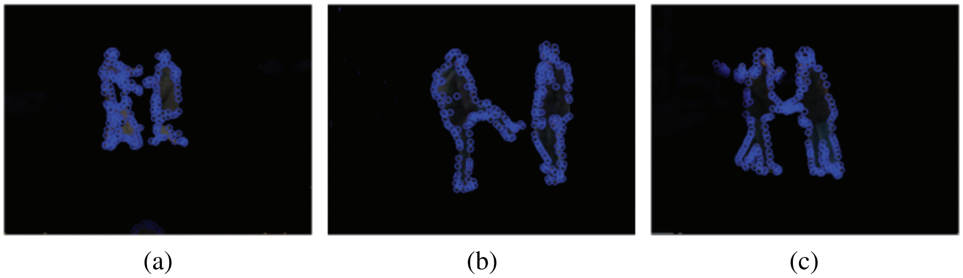


Figure 5: Features extraction using FAST, (a) show hug, (b) show kick, and (c) shake_hand

3.3.2 Oriented FAST and Rotated BRIEF

This study employed the FAST and ORB algorithm for feature extraction in the BIT Interaction and UT-Interaction dataset, which focuses on human-human interactions. ORB combines the features of the FAST keypoint detector and the ORB descriptor resulting in an efficient algorithm suitable for real-time applications. The algorithm first detects key points in the image using the FAST algorithm, which selects the points with a large difference in intensity for each neighboring pixel. Then it computes the orientation of each keypoint using the intensity distribution around it. Finally, the ORB descriptor is calculated for each keypoint. This binary string encodes the relative intensities of the pixel pairs around the keypoint. The resulting descriptors are robust to scale rotation and illumination changes, making them suitable for human-human interaction recognition.

Our experiments using the BIT Interaction dataset demonstrated the ORB algorithm's effectiveness in recognizing various human-human interactions. The ORB features were extracted from the images and used to train a machine-learning model, which achieved an accuracy of over 93%. And 94.6 using Ut-Interaction. The speed and accuracy of ORB make it an ideal feature extraction method for real-time applications such as human-human interaction recognition in video surveillance systems.

Computing the intensity centroid:

$$x_c = \frac{\sum_{x,y} I(x,y)\, x}{\sum_{x,y} I(x,y)}, y_c = \frac{\sum_{x,y} I(x,y)\, y}{\sum_{x,y} I(x,y)} \tag{11}$$

Computing the oriented ORB descriptor:

$$f_{\{d,i\}} = 1 \,\&\, if\ I(p_i) < I(p_j)\ \ 0 \,\&\, otherwise \tag{12}$$

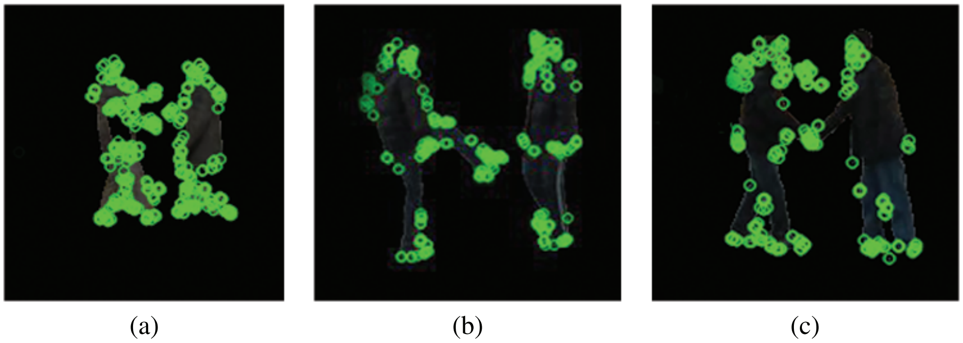
where p_i and p_j are pairs of points sampled from a circular region around the keypoint, and d denotes the descriptor index.

Computing the Hamming distance between two ORB descriptors:

$$d_H\left(f_1,f_2\right) = \sum_{i=1}^{N}\left(f_{1,i} \oplus f_{2,i}\right) \tag{13}$$

where N is the number of elements in the ORB descriptor, $\oplus$ denotes the bitwise XOR operation and $f_{1,i}$ and $f_{2,i}$ are the ith elements of the two descriptors being compared.

These equations are used in various stages of the ORB algorithm, such as keypoint detection descriptor computation and feature matching; results are also shown in Fig. 6.



(a) (b) (c)

Figure 6: Features extraction using ORB (a) show hug, (b) show kick, and (c) shake_hand

3.4 Feature Fusion and Discrimination

In recent years feature extraction has become integral to many computer vision applications, including object recognition, image matching, and scene reconstruction. One of the main challenges in feature extraction is achieving high accuracy and robustness, which requires a combination of multiple feature descriptors. In this paper, we explore the concept of feature fusion and discrimination for improving the performance of the BIT and UT interaction datasets. We extracted features using FAST and ORB feature detectors and saved them into a csv file. The next step is to fuse these features to create a more comprehensive dataset representation. To achieve this, we will explore several fusion methods, including feature concatenation, feature averaging, and feature weighting. Once the fused feature representation is obtained, we will use QDA to discriminate between different classes in the BIT interaction dataset. This approach can help overcome the limitations of using a single feature descriptor, improving performance and accuracy in computer vision tasks. Our experimental results demonstrate that the proposed feature fusion and discrimination approach outperforms the individual feature descriptors regarding discrimination accuracy, robustness, and speed.

Quadratic Discriminant Analysis (QDA) formula:

$$f\left(\mathrm{x}\right) = -\frac{1}{2}\left(\mathrm{x}-\mu\right)^{\mathrm{T}}\Sigma^{-1}\left(\mathrm{x}-\mu\right) - \frac{1}{2}\log\left|\Sigma\right| + \log \mathrm{P}\left(\mathrm{Y}=\mathrm{k}\right) \tag{14}$$

In this formula, $f(x)$ is the discriminant function that predicts the class of an observation x μ is the mean vector of the features Σ is the covariance matrix of the features $|\Sigma|$ denotes the determinant of Σ and $P\left(Y=k\right)$ is the prior probability of class k.

Weighted feature fusion formula:

$$\mathrm{F}_{\mathrm{fuse}} = \sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{w}_{\mathrm{i}} \mathrm{F}_{\mathrm{i}} \tag{15}$$

In this formula, F_{fuse} is the fused feature representation F_i are the individual feature descriptors w_i are the weights assigned to each feature descriptor and n is the total number of feature descriptors.

Feature discrimination formula:

$$d_{ij} = \frac{\left(\mu_{ij} - \mu_{ik}\right)}{\sqrt{\frac{1}{2}\left(\sigma_{ij}^2 + \sigma_{ik}^2\right)}} \tag{16}$$

where d_{ij} is the discriminant value of feature i for classes j and k, μ_{ij} is the mean of feature i in class j, μ_{ik} is the mean of feature i in class k and σ_{ij}^2 and σ_{ik}^2 are the variances of feature i in classes j and k, respectively. Fig. 7 represent features fusion and discrimination result using QDA.

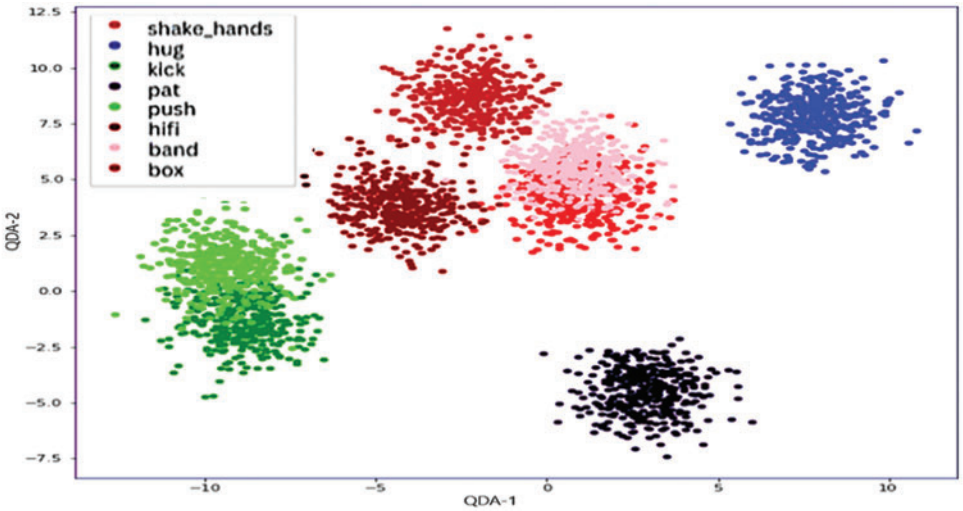


Figure 7: Features fusion and discrimination result using QDA

3.5 Hidden Markov Models

Hidden Markov Models (HMMs) are a class of probabilistic graphical models that capture the underlying dynamics of a system with hidden (unobservable) states [19–21]. These models have been widely used in speech recognition, natural language processing, bioinformatics and finance applications. In this research, we employ an HMM to model the hidden states and transitions of an 8-class dataset.

The following components define an HMM and also represent using Fig. 8:

A set of N hidden states, $S = \{s_1, s_2, \ldots, s_N\}$.

A set of M observable states, $O = \{o_1, o_2, \ldots, o_M\}$.

Transition probabilities between hidden states, $A = \left\{a_{\{ij\}}\right\}$, where $a_{\{ij\}} = P\left(s_{\{j\}}|s_{\{i\}}\right)$.

Emission probabilities of observable states given hidden states, $B = \left\{b_{\{j\}(k)}\right\}$, where $b_{\{j\}(k)} = P\left(o_k|s_j\right)$.

Initial state probabilities, $\pi = \{\pi_i\}$, where $\pi_i = P\left(s_i\right)$.

The HMM can be represented as a tuple $\lambda = (A, B, \pi)$.

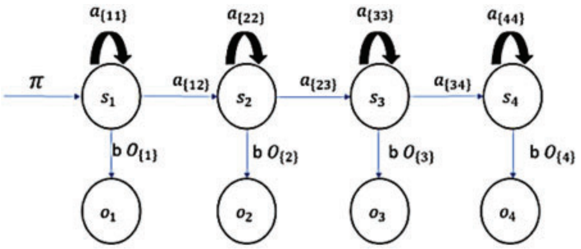


Figure 8: A simple illustration of a Hidden Markov Model. The circles represent hidden and observable states, while the arrows show the possible transitions between states

There are various methods to estimate the HMM parameters, including Maximum Likelihood Estimation (MLE) and the Expectation-Maximization (EM) algorithm, also known as the Baum-Welch algorithm. The MLE of the initial state probabilities π can be computed as:

$$\pi_i = \gamma_1 (i), \tag{17}$$

where $\gamma_1 (i)$ is the probability of being in state i at time 1, given the observations.

The MLE of the transition probabilities A can be computed as:

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t (i,j)}{\sum_{t=1}^{T-1} \gamma_t (i)}, \tag{18}$$

where $\xi_t (i,j)$ is the joint probability of being in states i and j at times t and t+1, respectively, given the observations, and $\gamma_t (i)$ is the probability of being in state i at time t given the observations.

The MLE of the emission probabilities B can be computed as:

$$b_j (k) = \frac{\sum_{t=1}^{T} \gamma_t (j)\, 1_{o_t=k}}{\sum_{t=1}^{T} \gamma_t (j)}, \tag{19}$$

where $1_{o_t=k}$ is an indicator function that is equal to 1 if $o_t = k$ and 0 otherwise.

Maximum Likelihood Estimation (MLE) of HMM parameters involves finding the parameters that maximize the likelihood of observing the given sequence of observations O. The likelihood of the observations can be expressed as:

$$P (O|\lambda) = \sum_S P (O, S|\lambda), \tag{20}$$

where $\lambda = (A, B, \pi)$ is the set of HMM parameters, and S is the set of possible hidden state sequences. The sum is taken over all possible state sequences that could have generated the observed sequence. Computing this sum directly is infeasible for large state spaces, but it can be solved efficiently using the Forward-Backward algorithm.

4 Experimental Analysis and Results

This study uses Hidden Markov Models (HMMs) as a classifier to analyze our proposed approach's performance. The experimental process was conducted with great attention to detail, and the numerical results were thoroughly analyzed. Our approach was evaluated using various measures such as precision, recall, F1-score, and support, which were all calculated to understand the classifier's performance comprehensively. The results showed that our HMM-based approach achieved an impressive accuracy of 93%, demonstrating the potential of our proposed method for real-world applications.

4.1 BIT/UT Interaction Dataset

BIT/UT [27] interaction dataset was used in this work. Additionally, our proposed solution was implemented in Visual Studio Code, and We took a dataset of human interaction frames and extracted features. The dataset was randomly divided into training and testing sets, with a 70% training size and a 30% testing size. We employed numerous measures to evaluate the performance of our suggested technique, including precision, recall, and F1-score, and we also provided support for each class in the dataset. BIT contains video recordings of human interactions from eight different classes: shake_hands, hug, kick, pat, push, hifi, bend, and box. The dataset is of exceptional quality with a resolution of 640 × 480 pixels and a total size of 4.4 GB. The videos were shot with a high-quality camera, showing various people engaging in natural interactions. The dataset was pre-processed to find out essential features for our HMM-based classifier. The resulting dataset was then used to test how well our proposed method worked.

4.2 Performance Measures

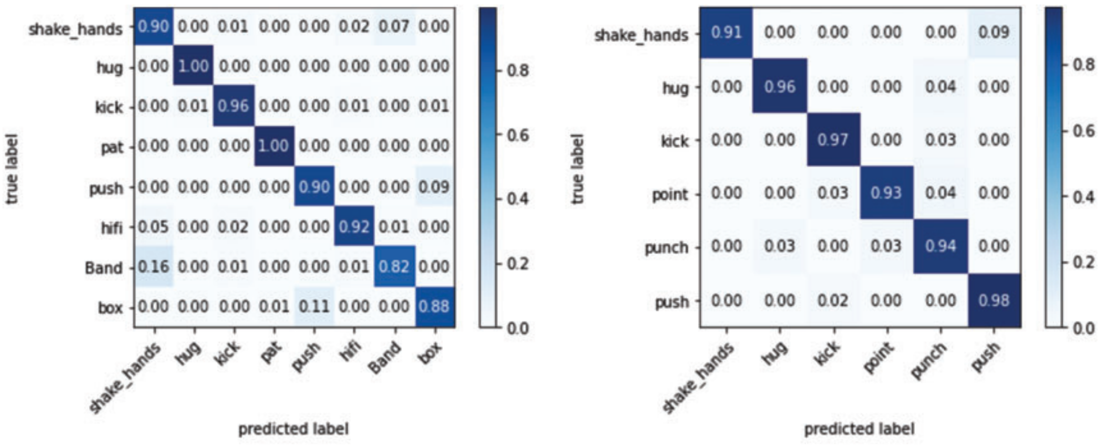
The result section provides an in-depth review of the recognition results generated with our suggested HMM-based technique. HMMs are a sort of probabilistic model that has been widely employed in pattern recognition and speech recognition tasks. HMMs represent a series of observations as hidden states that are not immediately observable but may be deduced from the observations. We used HMMs as a classifier in our suggested strategy to recognize human interactions based on a sequence of retrieved features. The HMMs were trained on the BIT/UT interaction dataset to classify new instances. HMMs have several advantages, including the capacity to simulate temporal dependencies and the flexibility to handle variable sequence lengths. These characteristics make HMMs well-suited for recognizing human interactions, which frequently entail complicated and diverse movement sequences. Our results show the efficacy and resilience of our recommended approach for detecting HHI in real-world contexts.

We tested our approach on the BIT interaction dataset, which performed admirably, achieving 93% accuracy overall. Support in machine learning refers to the number of occurrences in each dataset class. In Table 1 of our study, the "Support" column reflects the number of cases from each class utilized to train and evaluate our suggested approach. High recognition rates (F1-scores between 0.83) and reliability (F1-scores between 0.99) were attained using our method across all eight interaction classes. Specifically, our approach achieved a high recognition rate for Shake_hands (0.90), hug (1.00), kick (0.96), pat (1.00), push (0.90), hifi (0.92), bend (0.82), and box (0.88) interactions.

Fig. 9 shows the classification results from our HMM-based approach to recognizing human interactions. The matrix is essential for evaluating and improving classifiers. The confusion matrix shows that our method correctly classified most interaction types with only a few exceptions. Our method for identifying human interactions in real-world environments works well and is resilient. Table 2 represents a comparison of HIR accuracy using different techniques shown below.

Table 1: Performance measures of proposed HMM-based approach for recognizing human interactions

HIR-BIT-interaction					HIR-UT-interaction				
HIR	Precision	Recall	F1-score	Support	HIR-UT	Precision	Recall	F1-score	Support
SH	0.88	0.90	0.89	1601	**SH**	1.00	0.91	0.95	23
Hug	0.99	1.00	0.99	2904	**Punch**	0.92	0.94	0.93	127
Kick	0.96	0.96	0.96	1469	**Kick**	0.94	0.97	0.96	79
Pat	0.99	1.00	0.99	1976	**Push**	0.98	0.98	0.98	84
Push	0.87	0.90	0.89	1611	**Point**	0.97	0.93	0.95	134
Box	0.91	0.88	0.89	1949	**Hug**	0.95	0.96	0.95	72
Hifi	0.94	0.92	0.93	1004	**-**	–	–	–	–
Bend	0.85	0.82	0.83	890	–	–	–	–	–
Means	0.92	0.93	0.92	1675.5	**Means**	0.96	0.94	0.95	86.5



(a) (b)

Figure 9: Confusion, matrix of a proposed HMM-based approach for recognizing human interactions (a) displays BIT-Interaction results, while (b) displays UT-Interaction results

Table 2: Comparison of human interaction recognition accuracy using different techniques on the selected dataset

Techniques	SH	Hug	Kick	Pat	Punch	Hifi	Bend	Box	#Itration–time (s)	Detection rate (%)
CNN [23]	0.85	0.84	0.88	0.81	0.85	0.79	0.91	0.81	_	84.63
White stag model [29]	0.84	0.82	0.96	0.84	0.92	0.92	0.84	0.86	15–(22.5)	87.50
Two-stream [28]	0.87	1.00	1.00	0.75	0.81	0.87	1.00	0.93	_	90.63
Co-LSTM [30]	0.92	0.94	0.88	0.90	0.94	0.92	1.00	0.90	_	92.88
Proposed	0.90	1.00	0.96	1.00	0.90	0.92	0.82	0.88	15–(18.5)	93.00

5 Conclusion

This study presents a novel HMM-based method for recognizing human image interactions. This method has a high accuracy of 93% when applied to a dataset of human activities in BIT-Interaction and 94.6 using Ut-Interaction. The suggested approach includes several crucial phases: frame improvement and extraction, silhouette extraction, feature extraction, feature fusion and discrimination, and classification using HMM. Our method is also computationally effective, making it appropriate for real-time edge device applications. This study contributes to the expanding field of computer vision and pattern recognition and has real-world applications in biometrics, surveillance, and human-computer interaction. Future improvements can be made by incorporating deep learning techniques such as CNNs and RNNs for feature extraction and classification. Testing the proposed method on larger datasets and in more complex environments can also assess its generalizability and effectiveness.

Acknowledgement: None.

Funding Statement: The authors are thankful to the Deanship of Scientific Research at Najran University for funding this work under the Research Group Funding Program Grant Code (NU/RG/SERC/12/6). This study is supported via funding from Prince Satam bin Abdulaziz University Project Number (PSAU/2023/R/1444). Princess Nourah bint Abdulrahman University Researchers Supporting Project Number (PNURSP2023R348), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, and this work was also supported by the Ministry of Science and ICT (MSIT), South Korea, through the ICT Creative Consilience Program supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP) under Grant IITP-2023-2020-0-01821.

Author Contributions: Study conception and design: Tanvir Fatima Naik Bukht, Jaekwang Kim, data collection: Naif Al Mudawi, Saud S. Alotaibi and Aisha Ahmed AlArfaj; analysis and interpretation of results: Tanvir Fatima Naik Bukht, Abdulwahab Alazeb and Mohammed Alonazi; draft manuscript preparation: Tanvir Fatima Naik Bukht and Ahmad Jalal. 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.

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

References

[1] S. Nag, X. Zhu, Y. Z. Song and T. Xiang, "Proposal-free temporal action detection via global segmentation mask learning," in *European Conf. on Computer Vision*, Nature Switzerland, Tel Aviv, Israel, pp. 645–662, 2022.

[2] D. S. Sathiya, "Texture classification with modified rotation invariant local binary pattern and gradient boosting," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, vol. 26, no. 2, pp. 125–136, 2022.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, vol. 1, pp. 886–893, 2005.

[4] H. Rahman, T. F. N. Bukht, A. Imran, J. Tariq, S. Tu *et al.,* "A deep learning approach for liver and tumor segmentation in CT Images using ResUNet," *Bioengineering*, vol. 9, no. 8, pp. 1–9, 2022.

[5] A. Nadeem, A. Jalal and K. Kim, "Human actions tracking and recognition based on body parts detection via artificial neural network," in *IEEE Int. Conf. on Advancements in Computational Sciences*, Lahore, Pakistan, pp. 1–6, 2020.

[6] H. Rahman, T. F. N. Bukht, R. Ahmad, A. Almadhor, A. R. Javed *et al.,* "Efficient breast cancer diagnosis from complex mammographic images using deep convolutional neural network," *Computational Intelligence and Neuroscience*, vol. 2023, pp. 1–11, 2023.

[7] P. U. Putra, K. Shima and K. Shimatani, "A deep neural network model for multi-view human activity recognition," *Public Library of Science*, vol. 17, no. 1, pp. 1–20, 2022.

[8] O. M. Sincan and H. Y. Keles, "Using motion history images with 3D convolutional networks in isolated sign language recognition," *IEEE Access*, vol. 10, pp. 18608–18618, 2022.

[9] T. F. N. Bukht, H. Rahman and A. Jalal, "A novel framework for human action recognition based on features fusion and decision tree," in *2023 4th Int. Conf. on Advancements in Computational Sciences (ICACS)*, Lahore, Pakistan, pp. 1–6, 2023.

[10] M. Quaid and A. Jalal, "Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm," *Multimedia Tools and Applications*, vol. 79, pp. 6061–6083, 2020.

[11] M. Batool, A. Jalal and K. Kim, "Sensors technologies for human activity analysis based on SVM optimized by PSO algorithm," in *2019 Int. Conf. on Applied and Engineering Mathematics (ICAEM)*, Taxila, Pakistan, pp. 145–150, 2019.

[12] M. A. Khan, M. Sharif, T. Akram, M. Raza, T. Saba *et al.,* "Handcrafted and deep convolutional neural network features fusion and selection strategy: An application to intelligent human action recognition," *Applied Soft Computing*, vol. 87, pp. 1–30, 2020.

[13] A. Mounsey, K. Asiya and S. Sanjay, "Deep and transfer learning approaches for pedestrian identification and classification in autonomous vehicles," *Electronics*, vol. 10, no. 24, pp. 3159–3175, 2021.

[14] N. Y. Khattak, E. G. Hussnain, W. Ahmad, S. U. Islam and I. U. Haq, "Computer vision-based human activity classification and prediction," *The Sciencetech*, vol. 4, no. 1, pp. 45–48, 2023.

[15] M. Li, T. Li and Y. Jiang, "Marker displacement method used in vision-based tactile sensors from 2D to 3D—A review," *IEEE Sensing Journal*, vol. 23, no. 8, pp. 8042–8059, 2023.

[16] T. Luczak, P. Nelsen, J. E.Ball, R. Burch, J. Barlow *et al.,* "A survey of technical challenges in computer vision via machine and deep learning for human pose estimation," in *IIE Annual Conf. Proc.*, Norcross, Georgia, pp. 1–6, 2022.

[17] A. Jalal, S. Kamal and D. Kim, "Depth silhouettes context: A new robust feature for human tracking and activity recognition based on embedded HMMs," in *2015 12th Int. Conf. on Ubiquitous Robots and Ambient Intelligence (URAI)*, Goyangi, Korea, pp. 294–299, 2015.

[18] L. Jia, X. Zhou and C. Xue, "Non-trajectory-based gesture recognition in human-computer interaction based on hand skeleton data," *Multimedia Tools and Applications*, vol. 81, no. 15, pp. 20509–20539, 2022.

[19] L. Jin, J. Cheng and C. Zhang, "Infrared pedestrian tracking with graph memory features," *IEEE Signal Processing Letters*, vol. 28, pp. 1933–1937, 2021.

[20] Y. Du, W. Wang and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 1110–1118, 2015.

[21] H. Ansar, A. Jalal, M. Gochoo and K. Kim, "Hand gesture recognition based on auto-landmark localization and reweighted genetic algorithm for healthcare muscle activities," *Sustainability*, vol. 13, no. 5, pp. 1–26, 2021.

[22] S. A. Rizwan, A. Jalal, M. Gochoo and K. Kim, "Robust active shape model via hierarchical feature extraction with SFS-optimized convolution neural network for invariant human age classification," *Electronics*, vol. 10, no. 4, pp. 1–24, 2021.

[23] A. Jalal, M. Mahmood and A. S. Hasan, "Multi-features descriptors for human activity tracking and recognition in indoor-outdoor environments," in *2019 16th Int. Bhurban Conf. on Applied Sciences and Technology (IBCAST)*, Islamabad, Pakistan, pp. 371–376, 2019.

[24] N. Khalid, M. Gochoo, A. Jalal and K. Kim, "Modeling two-person segmentation and locomotion for stereoscopic action identification: A sustainable video surveillance system," *Sustainability*, vol. 13, no. 2, pp. 1–30, 2021.

[25] M. Teutsch, A. Sappa and R. Hammoud, "Computer vision in the infrared spectrum: Challenges and approaches," *Synthesis Lectures on Computer Vision*, vol. 10, no. 2, pp. 1–138, 2021.

[26] A. Jalal and M. Mahmood, "Students' behavior mining in e-learning environment using cognitive processes with information technologies," *Education and Information Technologies*, vol. 24, no. 5, pp. 2797–2821, 2019.

[27] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *IEEE Int. Conf. on Computer Vision (ICCV)*, Kyoto, Japan, pp. 1593–1600, 2009.

[28] J. Zhou, Z. Wang, J. Meng, S. Liu, J. Zhang *et al.,* "Human interaction recognition with skeletal attention and shift graph convolution," in *2022 Int. Joint Conf. on Neural Networks (IJCNN)*, Padua, Italy, pp. 1–8, 2022.

[29] M. Maria, A. Jalal and K. Kim, "WHITE STAG model: Wise human interaction tracking and estimation (WHITE) using spatio-temporal and angular-geometric (STAG) descriptors," *Multimedia Tools and Applications*, vol. 79, no. 11, pp. 6919–6950, 2020.

[30] X. Shu, J. Tang, G. J. Qi, Y. Song, Z. Li *et al.,* "Concurrence-aware long short-term sub-memories for person-person action recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Nanjing, China, pp. 1–8, 2017.