Computers, Materials & Continua
DOI: 10.32604/cmc.2023.033769
Article

Exploiting Human Pose and Scene Information for Interaction Detection

Manahil Waheed[1], Samia Allaoua Chelloug[2,*], Mohammad Shorfuzzaman[3], Abdulmajeed Alsufyani[3], Ahmad Jalal[1], Khaled Alnowaiser[4] and Jeongmin Park[5]

[1]Department of Computer Science, Air University, Islamabad, 44000, Pakistan
[2]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia
[3]Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, 21944, Saudi Arabia
[4]Department of Computer Engineering, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, 11942, Saudi Arabia
[5]Department of Computer Engineering, Korea Polytechnic University, Siheung-si, Gyeonggi-do, 237, Korea
*Corresponding Author: Samia Allaoua Chelloug. Email: sachelloug@pnu.edu.sa
Received: 27 June 2022; Accepted: 22 September 2022

Abstract: Identifying human actions and interactions finds its use in many areas, such as security, surveillance, assisted living, patient monitoring, rehabilitation, sports, and e-learning. This wide range of applications has attracted many researchers to this field. Inspired by the existing recognition systems, this paper proposes a new and efficient human-object interaction recognition (HOIR) model which is based on modeling human pose and scene feature information. There are different aspects involved in an interaction, including the humans, the objects, the various body parts of the human, and the background scene. The main objectives of this research include critically examining the importance of all these elements in determining the interaction, estimating human pose through image foresting transform (IFT), and detecting the performed interactions based on an optimized multi-feature vector. The proposed methodology has six main phases. The first phase involves preprocessing the images. During preprocessing stages, the videos are converted into image frames. Then their contrast is adjusted, and noise is removed. In the second phase, the human-object pair is detected and extracted from each image frame. The third phase involves the identification of key body parts of the detected humans using IFT. The fourth phase relates to three different kinds of feature extraction techniques. Then these features are combined and optimized during the fifth phase. The optimized vector is used to classify the interactions in the last phase. The MSR Daily Activity 3D dataset has been used to test this model and to prove its efficiency. The proposed system obtains an average accuracy of 91.7% on this dataset.

Keywords: Artificial intelligence; daily activities; human interactions; human pose information; image foresting transform; scene feature information

1 Introduction

Human interaction recognition (HIR) is an emerging field in computer vision and artificial intelligence. It involves identifying humans and the objects they are interacting with and then mining various features to help the classifier identify the correct interaction. Such interaction recognition systems are being widely used in the health [1–3], sports [4,5], security [6], and education [7] sectors. This article uses daily activities to test the proposed model. In daily lives, humans interact with various objects; for example, they drink water from a glass or cup, eat from a plate or packet, read or write on a paper, play an instrument, or use a device [8]. If an automatic system is able to identify these interactions of a patient, it can help provide them with timely assistance [9]. Similarly, if the interactions of a student are monitored during an online class, it can enhance their learning experience [10]. Moreover, if such interactions are identified at a certain place, better security can be provided [11].

Despite the various uses mentioned above and the excessive amount of research done in this field, there is still a lot of room for improvement. The existing systems are not a hundred percent accurate and don't work as well in practice. One of the main reasons for this is the numerous challenges that a recognition system might face. These challenges range from occlusion, scale variation, inter-class similarities, intra-class differences, and illumination variations [12]. Another major problem is the limited availability of large-scale publicly available datasets that provide all these issues so a system can be trained for real-life challenges. Keeping all these in mind, we have proposed a novel method for interaction recognition in this paper.

Some existing HOIR systems rely on scene information as contextual clues provide extra information that helps in determining the HOI [13]. Others remove background scenes and consider only the foreground [14]. The rationale behind this approach is that redundant background negatively contributes to the process of deciding the correct interaction. The remaining models use interaction points [15] or a combination of full humans and their key points to identify the interaction being performed [16]. Our approach uses a hybrid of these three approaches as we try to establish the importance of all kinds of clues. Scene information is important because if a person is holding something on a cricket ground, that something is more likely to be a ball or a bat, but if it is a kitchen scene, the object is probably a kitchen utensil. Similarly, pose estimation is important because some key body parts are more involved in a certain interaction than others. This will make the human pose for such an interaction distinctive.

Our proposed model includes six main steps. The first step involves preprocessing the incoming videos containing human interactions. First, the videos are converted into frames and then these frames are adjusted and refined using nose removal. In the second step, the human-object pair is detected through image segmentation. Then the twelve key body parts of the detected humans are identified during the third step. The fourth step consists of various feature extraction techniques. The different types of features are extracted. First, full human features are obtained, and then features based on key body parts are extracted. Lastly, scene features are obtained. In the fifth step, these features are concatenated, and the resulting vector is optimized. This vector is then fed to the hidden Markov model (HMM), which recognizes the interaction. The main contributions of this paper are as follows:

- Using a graph-based image skeletonization technique called IFT for pose estimation that involves detecting twelve human body parts.
- Proposing a multi-feature approach involving three different types of features: full-body features, point-based features, and scene features.
- Optimizing the large feature vector obtained through locally linear embedding (LLE).

- Using HMM for the final class detection of daily-life activities involving human-object interactions.

The rest of this research article is arranged as follows: Section 2 discusses the related research work to some extent. Section 3 explains the proposed method and its various phases in detail. Then the proposed model is evaluated in Section 4. The dataset and the experimental settings and results are discussed. Section 5 touches upon some limitations of this model and possible future directions of research that can improve this existing model. Finally, Section 6 provides a conclusion of this research.

2 Literature Review

In this article, an efficient human interaction recognition (HIR) system has been proposed which uses human pose and scene feature information to accurately identify the interactions. Many previous researchers have employed human pose for this purpose and others have used scene information, which has more commonly been referred to as contextual information. These research works are discussed in detail below:

2.1 HIR Based on Human Pose

Modeling human pose is important because different human poses and the movement of body parts are related to different interactions. Hence, many research articles have given methods of detecting body parts and extracting features from them for better human-object interaction (HOI) classification. Yao et al. [17] exploited the mutual context of the overall human pose, different body parts, and the object involved in an interaction. They argue that the two difficult tasks of object detection and human pose estimation can benefit from each other by providing mutual context. Similarly, Ghadi et al. [18] proposed a model for detecting twelve human body parts using binary silhouettes. The authors then extracted features from full humans as well as key body parts to obtain a rich and robust feature vector. For full humans, they used oriented FAST and rotated BRIEF (ORB) and texton map features. And for body parts, they obtained Radon transforms and Freeman chain codes. A similar method was followed by Khalid et al. [19] for HOI detection. They use 3 dimensional (3-D) point clouds and fiducial points as features. They also exploited the overall pose and body parts information for better classification. Similarly, Waheed et al. [20] extracted six key body points using heat kernel signatures of 3-D meshes and geodesic distances. Moreover, they extracted convolutional neural network (CNN)-based features for human pose and topological and geometric features for key human body parts. Likewise, Wang et al. [21] proposed the use of interaction points for HOI detection. The authors employed a fully convolutional approach that directly detected interaction points between the human and the object. Based on these points, their model generated an interaction vector. However, all these systems are limited to humans and their body parts. These do not utilize contextual information.

2.2 HIR Based on Scene Information

Scene features are important because they provide additional contextual information about the interactions. For example, if the scene features identify the scene as a playground, it is easier to detect that the performed interaction is a sports activity. Many previous works have exploited this additional information to improve the accuracy of their systems. For example, Ikizler-Cinbis et al. [22] used two types of scene features for human interaction recognition; namely, shape features and color features. They used GIST as shape features and for color features, they divided the input frames into three regions and obtained their histograms. Similarly, Wang et al. [23] employed deep contextual features for human-object interaction recognition. For this, they obtained CNN features of whole images.

They called them global features and then also obtained local features from the detected human and object instances. He et al. [24] used scene graphs for HOIR. They detected the human and object pairs and then added the external environment or scene information to provide contextual clues. The scene graphs also provided spatial clues, such as the size and position of the object. Another interesting approach was presented by Gupta et al. [25], who treated the daunting task of human-object interaction recognition as a Bayesian problem. To establish the importance of using contextual clues, the authors give an example of a simple running interaction that will more likely be interpreted as kicking if there is a ball right next to it. The issue with these models is that they rely too heavily on scene information and too little on the more important human pose.

3 Material and Methods

The proposed system has six different stages. First, frames are extracted from the input videos and then the images are preprocessed using intensity adjustment and noise removal techniques. In the second stage, image segmentation is performed where each image is converted into super pixels and then these super pixels are merged together to get the desired human-object (HO) pairs. Next, the key human body parts are identified on the detected HO pairs during the third stage. In the fourth stage, three types of features are then obtained: full body, parts-based, and scene features. In the fifth stage, the three different features are concatenated and optimized. The interactions are then classified in the sixth stage. A complete overview of the proposed model is given in Fig. 1.

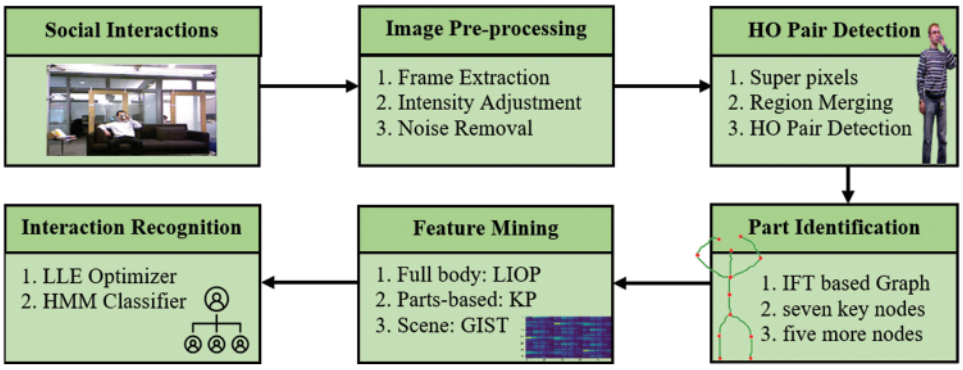


Figure 1: A complete overview of the proposed model

3.1 Image Preprocessing

All input videos are first converted into frames at the rate of 20 frames per second. Then these frames are pre-processed. The intensity values of the images are adjusted using sigmoid stretching and then noise is removed using Gaussian filtering. These two methods are discussed in detail in the following sub-sections.

3.1.1 Sigmoid Stretching

For adjusting the contrast of all images, we use sigmoid stretching, which is a linear image transformation method based on piecewise linear functions. It enhances the quality of an image by improving the contrast. The intensity values are stretched to fill the entire dynamic range of an image. For this, a linear and monotonically increasing transformation function is used. This method highlights the pixels with moderate intensity values and maintains enough contrast at the extreme pixels. Therefore, all the pixel values seem to be placed along a sigmoidal function (an S-shaped curve). Hence, the resultant image has less contrast in very bright and very dark areas, and more contrast in

areas between these two extremes. This kind of stretching works ideally for almost every kind of image. Eq. (1) shows the sigmoid function.

$$Sigmoid\ (x) = \frac{1}{(1 + e^{-x})} \tag{1}$$

3.1.2 Gaussian Filtering

After intensity adjustment, noise is removed from the images. For this, we use Gaussian filtering. A gaussian filter has an impulse response of a Gaussian function. It works as a 2-D convolution operator which removes noise from the image by making it smooth. However, some of the details are also lost during the process and hence, the resultant image seems blurred. In the output image, each pixel is a weighted average of the neighbors of the pixel in the original image. The average is weighted more towards the value of the central pixels. It is quite similar to a mean filter, but it uses a different kernel. This kernel represents a bell-shaped Gaussian hump, as given in Eq. (2).

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

3.2 Human-Object Pair Detection

After preprocessing, all images are segmented. This means that the images are divided into background and foreground segments, where the foreground segment contains the human-object (HO) pair. For this purpose, we employed Felzenszwalb's method [26], which divides the input image into multiple regions called super pixels. Felzenszwalb's method creates a graphical representation of the input image and then decides which areas are similar and can be categorized as one region. Fig. 2 shows the results of this method. Then these regions are merged further until only 3 regions are left: the background, the human, and the object. This region merging technique is inspired by the work of Xu et al. [27], according to which, the regions which are similar and adjacent are merged to get one bigger region. To determine the similarity, four types of features are obtained for each region. These features include mean, covariance, scale-invariant feature transform (SIFT), and speeded-up robust features (SURF). Eq. (3) shows the similarity $S_{i,j}$ formula.

$$S_{i,j} \leftarrow \sum_{i=1,j=1}^{n} \left[S_{i,j}^{mean} + S_{i,j}^{covariance} + S_{i,j}^{sift} + S_{i,j}^{surf}\right] \times D_{i,j} \tag{3}$$

where i and j represent any two regions or super pixels, $D_{i,j}$ is their adjacency matrix. This means that it is equal to 1 if the two regions are adjacent and 0 otherwise. $S_{i,j}^{mean}$, $S_{i,j}^{covariance}$, $S_{i,j}^{sift}$ and $S_{i,j}^{surf}$ are the similarities between the mean, covariance, SIFT and SURF features of the two regions.

3.3 Part Identification

After extracting the human silhouette, twelve key human body parts have been identified. For this purpose, the first step is to convert the human silhouette into a binary silhouette whose image skeleton is then obtained. In the binary image, the foreground is black and the background is white. The process of skeletonization keeps reducing the foreground until no more pixels can be removed. The skeletal remnant is then used to identify key points. For image skeletonization, a graph-based technique called the image foresting transform (IFT) [28] has been used. IFT represents the path in a graph with the minimum cost. In such a path forest, the nodes represent the pixels of the input image and the adjacency relationship between various pixels represents its arcs. To determine the cost of a path in such a graph, a cost function specific to this application is calculated. Moreover, the IFT depends on the local properties of the image along the path. For example, the color, gradient, or the

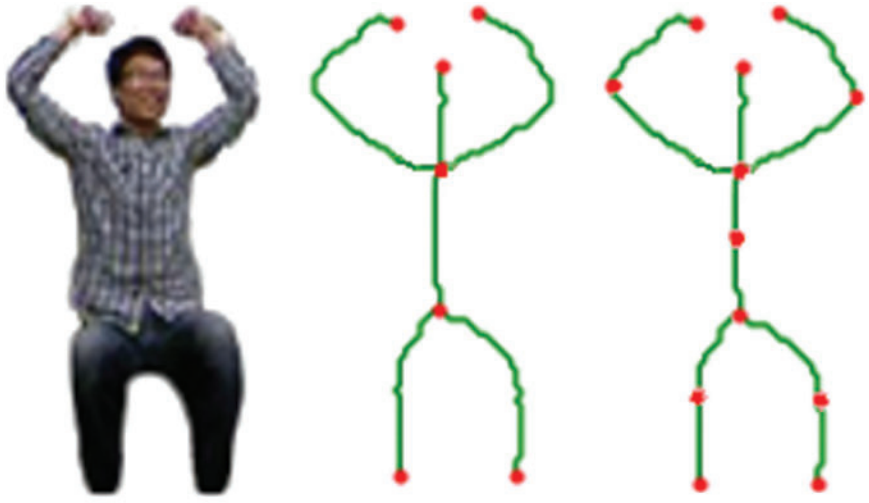
position of the pixel. A given set of seed pixels is used to determine the roots of this path forest. To get suitable cost functions for the different paths, one minimum-cost path is assigned to each pixel from the given set of seed pixels. This is done so that the paths are connected to form an oriented forest that spans the entire image. There are the outputs of an IFT: an optimum path from the root, the path cost, and the corresponding root. The path attribute can be used to find an image skeleton.

(a) (b) (c) (d)

Figure 2: HO pair detection (a) original image, (b) super-pixels, (c) regions merged and (d) detected HO pair

7 key points are obtained from the nodes marking the start and end positions of various paths in the obtained skeleton. These points are identified as head, left hand, right hand, upper torso, bottom torso, left foot and right foot. Using the obtained 7 points, 5 additional key points are also found, namely, neck, left elbow, right elbow, left knee, and right knee. The method of finding these additional points is simple: the mid-point of any two key points is calculated and a point on the contour lying closest to the obtained mid-point is stored as an additional key point. The mid-point (x_m, y_m) of two existing points j and k is calculated using Eq. (4). Each step of the process is shown in Fig. 3.

$$(x_m, y_m) = \left(\frac{x_j + x_k}{2}, \frac{y_j + y_k}{2}\right) \tag{4}$$



(a) (b) (c)

Figure 3: Part identification (a) overall human pose (b) seven key IFT graph nodes and (c) five additional nodes

3.4 Feature Extraction

Robust features play a critical role in identifying an HOI interaction. Therefore, three different types of features have been employed by the proposed system: full body, parts-based, and scene features. Algorithm 1 explains this process in detail while each type of feature is discussed in the following sub-sections.

Algorithm 1: Feature Extraction

```
Input: N: full body silhouettes and twelve key body points
Output: combined feature vector f = (f1, f2, f3 . . . fn)
% initiating feature vector for remote sensing HOI classification %
feature_vector ← []
% loop on all images%
n ← len (images)
For i = 1:n
    % extracting scene features%
    GIST ← Get_GIST_descriptor (image[i])
    feature_vector.append (GIST)
    % loop on extracted human silhouettes %
    s ← len (silhouettes)
    For j = 1:s
        % extracting intensity order pattern (LIOP) features%
        LIOP ← GetLIOPdescriptor (silhouette[j])
        feature_vector.append (LIOP)
        % loop on twelve key points %
        For k = 1:12
            % extracting kinematic posture (KP) features%
            KP ← GetKinematicPosture (k, k + 1)
            feature_vector append (KP)
        End
    End
End
Feature-vector ← Normalize (feature_vector)
return Feature-vector f = (f1, f2, f3 . . . fn)
```

3.4.1 Full Body: LIOP Feature

For the overall human pose or full body silhouette of the detected human, we have extracted the local intensity order pattern (LIOP) [29]. This feature descriptor works better in the case of low contrast and illumination changes within an image but faces issues in the case of rotation and scale variation. It is also useful with challenges such as geometric and photometric transformations. These can include view-point changes, blur images, and compressed images. The LIOP feature descriptor is based on the order of intensity values. To obtain this feature, the image region containing the detected HO pair is divided into multiple sub-regions called ordinal bins based on their intensity order. Then the LIOP descriptor of each point in the sub-region is obtained on the basis of its relationship with the intensities of its neighboring points. If $P(x)$ is a vector that contains the intensity values of the neighbors of a point that belongs to the local patch, then Eq. (5) can be used to get the LIOP of

this point. In this equation, $\omega(x)$ denotes a weight function which is described in Eq. (6). The LIOP descriptor is obtained by concatenating the LIOPs of the various points of each bin, respectively. Each image is divided into 6 bins and the number of neighboring points is set to 4, resulting in a feature vector of size $4! \times 6 = 144$. Each step of this process is shown in Fig. 4, and Eq. (7) represents the LIOP descriptor of each bin.

$$\omega(x) = \sum_{i,j} sgn\left(\left|I(x_i) - I\left(x_j\right)\right| - T_{lp}\right) + 1 \tag{5}$$

$$LIOP(x) = \varnothing(\gamma(P(x))) \tag{6}$$

$$des_i = \sum_{x \in bin_i} \omega(x)\, LIOP(x) \tag{7}$$

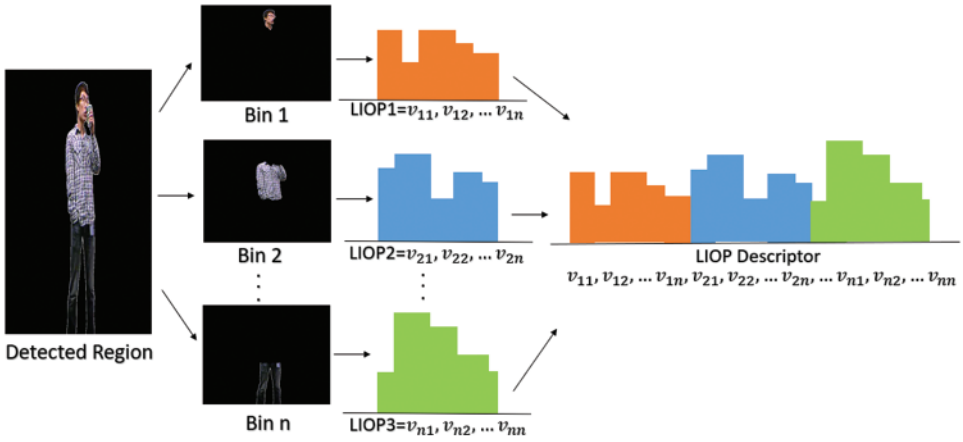


Figure 4: LIOP feature descriptor

3.4.2 Parts-Based: Kinematic Posture

For body parts, we have obtained a feature known as kinematic posture [30]. It includes two feature sets: linear joint position feature (LJPF) and angular joint position feature (AJPF). To obtain this feature, every key body part i is represented by a three-dimensional vector J_i in the coordinate space of Kinect. Then the distance of each body part is obtained with respect to the head J_{head}. This distance $d_{i(head)}$ is then normalized with respect to the distance vector between neck and torso points. Hence, for twelve key points, the LJPF for each frame n can be represented by Eq. (8).

$$LJPF_n = \left[d_{[n,1]}, d_{[n,2]}, \ldots d_{[n,12]}\right] \tag{8}$$

Then the angles between different bone segments are calculated using three body parts. The AJPF encodes the angles between different bone segments. For example, the angle between the left upper arm and forearm is calculated using the neck, left elbow, and left-hand join. Since the angle between the neck and the head is almost constant for all actions, only five angles are computed. Hence, the AJPF for each frame n can be represented by Eq. (9).

$$AJPF_n = \left[a_{[n,1]}, a_{[n,2]}, \ldots a_{[n,5]}\right] \tag{9}$$

Lastly, for each video frame, these two features are combined to generate the kinematics posture feature (KPF) set. This feature encodes the change in key body positions and angles across video frames. Both LJPF and AJPF features of an image skeleton are shown in Fig. 5.

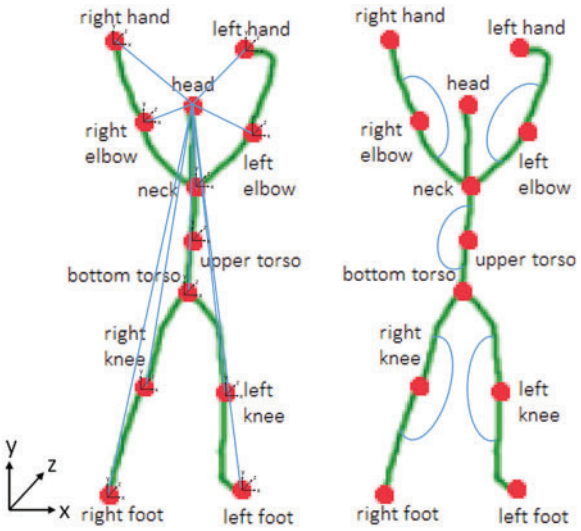


Figure 5: Kinematic posture (a) LJPF (b) AJPF

3.4.3 Scene: GIST Descriptor

GIST [31] feature descriptor is a global texture-based feature extraction technique that is used for extracting the dominant structure of a scene. This feature representation is usually based on five perceptual dimensions, i.e., roughness, ruggedness, naturalness, expansion, and openness. Initially, the input frames are converted into gray-scale images. To obtain the GIST descriptor, the input frame is first convolved with 32 Gabor filters. These filters have 4 scales (σ) and 8 orientations (θ), resulting in a series of 32 feature maps. These maps have the same size as the input image frame. Each feature map is divided into 9 regions and then the values within each region are averaged. These 9 values of the 32 feature maps are then joined together to give the 288-dimensional GIST feature vector for each frame. The GIST descriptors of two different scenes are visualized in Fig. 6.

3.5 Feature Optimization

Once the three types of features are obtained, they are concatenated. However, this results in a high-dimensional feature vector in which each dimension represents a specific feature. To reduce the dimensions of this feature vector and to make the system computational effective, we use a technique called locally linear embedding (LLE) [32]. It is an unsupervised dimensionality reduction method for non-linear data. It maps the high-dimensional data to lower dimensions while preserving its neighborhood embeddings. It does so by representing each point in the original data as a regularized linear mixture of its neighboring points. As explained in Algorithm 2, LLE works on each point by creating its neighborhood graph using its K-ary neighborhoods.

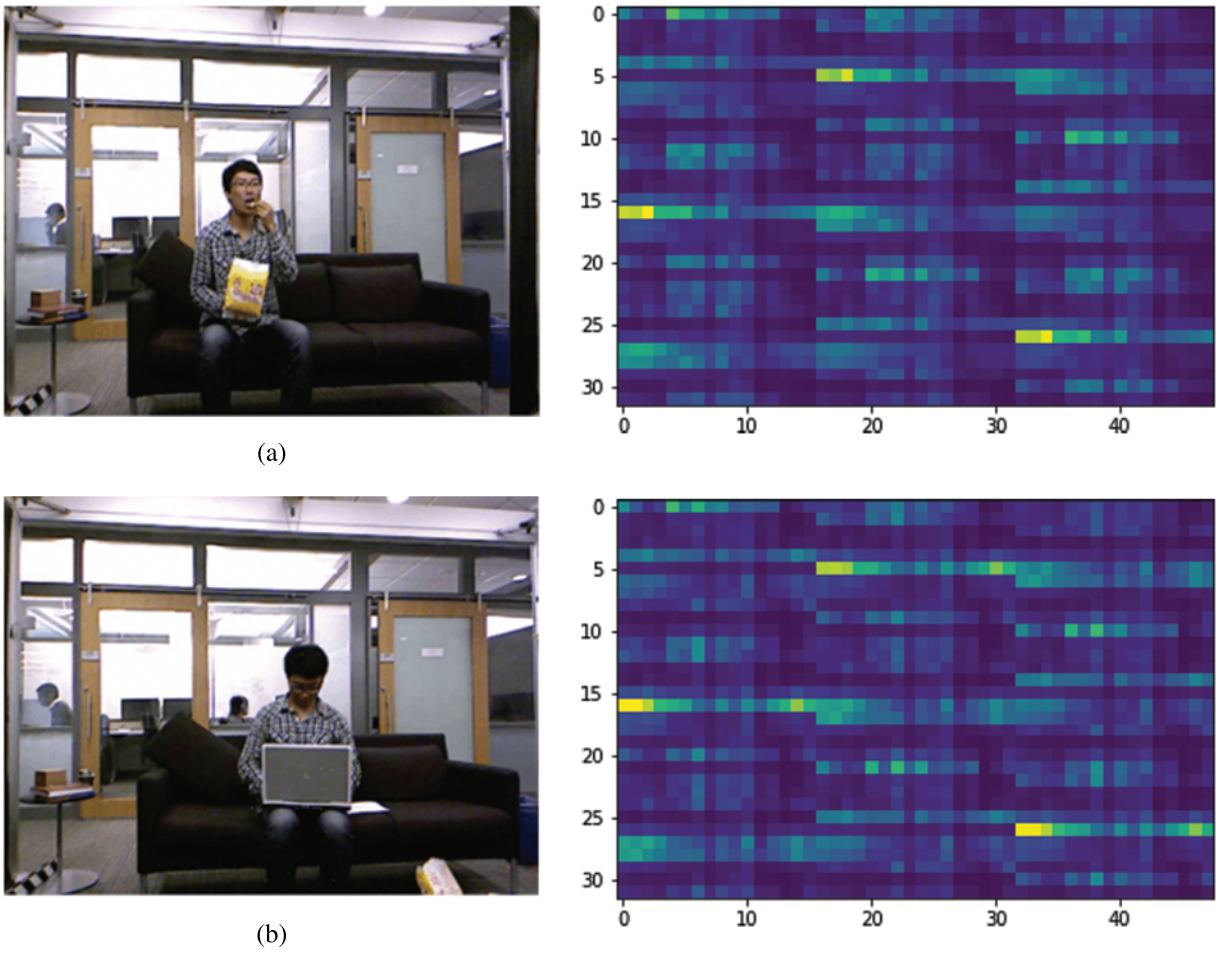


Figure 6: (a) GIST descriptor (a) eat (left) and its GIST descriptor (right) (b) use laptop and its GIST descriptor

Algorithm 2: Feature Optimization

Input: high dimensional feature vector f
Output: low dimensional feature vector Y
%find neighbors of each data point x%
For i ← 1:N
    Distance_vector ← *[ ]*
    For j ← 1:N
        D ← *distance* $(x_i x_j)$
        Distance_vector.append (D)
    END
    d ← Get_K_smallest_distances (D)
    Z ← *neighbors_of_x(d)*
END

(Continued)

Algorithm 2: Continued

%solving for reconstruction weights W%
For i = 1:N
    %compute local covariance%
    C ← Z′ ∗ Z [e]
    %solve linear system%
    C∗w = 1 for w [f]
        If j is not a neighbor of i:
            Wij ← 0
        Else:
            W ← w/sum(w)
END
%computing the output vector Y%
%Create sparse matrix M%
M ← (I − W)′ ∗ (I − W)
E ← Find_bottom_d + 1_eigenvectors (M)
Yq ← q + 1_smallest_eigenvectors (E)
Return Y

3.6 Interaction Classification

For the classification of interactions, a hidden Markov model (HMM) [33] is used. It is a probabilistic machine learning model that uses some observed or known variables to predict a sequence of hidden or unknown variables. In this method, system X is assumed to have unobservable or hidden states, but these states can be determined through another observable process Y. Although 1-D HMMs are more popular, 2-D HMMs are used for image classification. During training, the model parameters are estimated based on the obtained feature vectors and their labeled classes. During testing, the trained HMM searches those classes which have the maximum a posteriori probability given the feature vectors. Fig. 7 shows the structure of a 2-D HMM.

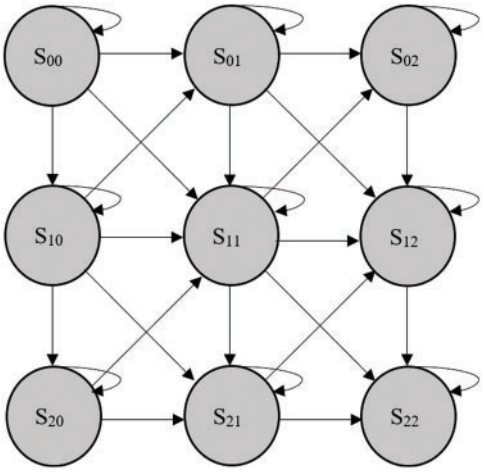


Figure 7: 2-D HMM structure

4 Experiments and Results

This section describes the dataset that has been used for experimentation and the details of the various experiments performed. The leave one subject out (LOSO) cross-validation technique has been used to evaluate the proposed model. First, classification accuracy is given in terms of the confusion matrix. It shows the accuracy achieved with each class. Moreover, precision, sensitivity, specificity, and F1-score values are also given. Then body part detection rate for each part is given by comparing the obtained coordinates with the ground truth values. The entire training and testing process was carried out using Python on a Windows-10 operating system which had 16-GB RAM and a core-i7-7500U CPU @ 2.70 GHz processor. Lastly, the performance of the proposed system is also compared with that of other state-of-the-art models that used the same dataset.

4.1 MSR Daily Activity 3D Dataset

The MSR Daily Activity 3D dataset [34] contains both depth and RGB video sequences, which have been recorded using a Kinect sensor at Microsoft Research Redmond. 10 different people perform 16 activities, only 10 of which are human-object interactions, namely, *read book, drink, call cellphone, eat, use vacuum cleaner, write on a paper, toss paper, play game, use laptop,* and *play guitar*. Every person performs the given interaction once while standing and once while sitting. The dataset has a total of 320 videos, but we have used only 200 videos of the 10 human-object interactions. This dataset is challenging because of high intra-class variation. Some sample frames from the dataset are given in Fig. 8.

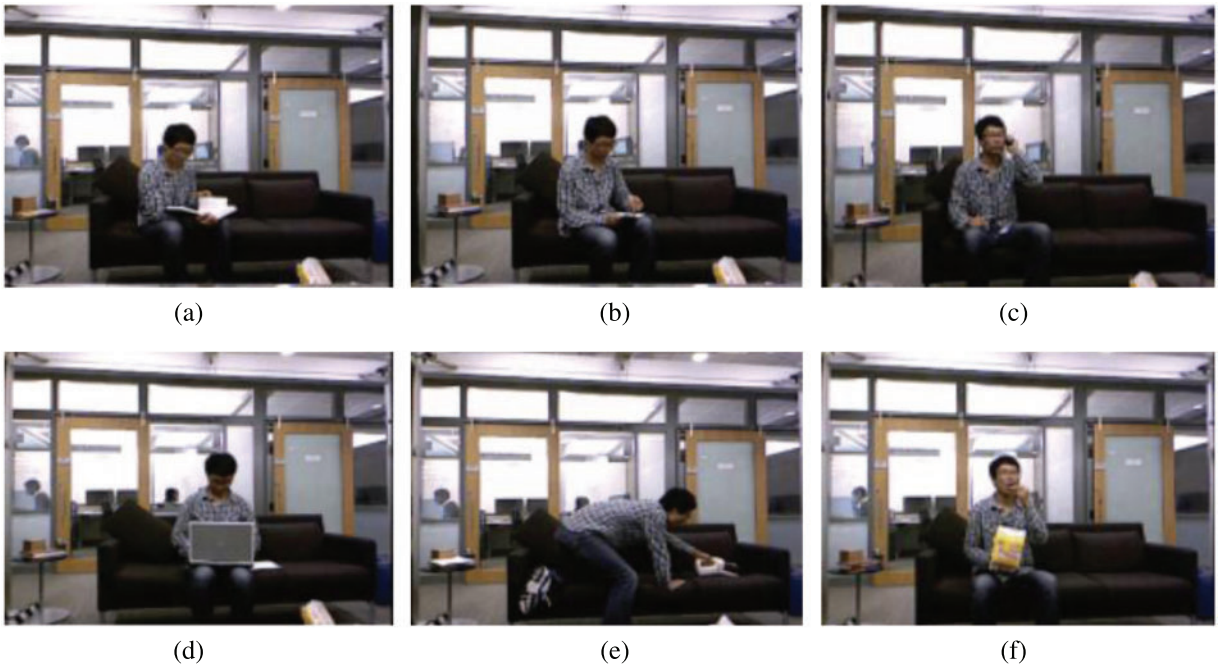


Figure 8: MSR daily activity 3D dataset. a) read book, b) write on a paper, c) call cellphone, d) use laptop, (e) use vacuum cleaner and (f) eat

4.2 Classification Accuracy

The proposed model has achieved an average accuracy of 91.7% on the MSR Daily Activity 3D dataset and the results are given below. Table 1 shows a confusion matrix of the proposed model. It shows the number of times (in terms of percentages) each class was identified correctly as well as when it was incorrectly identified as another class.

Table 1: Confusion matrix over the MSR Daily Activity dataset

Classes	DR	ET	RB	WP	UL	PG	CC	UV	PR	LS
DR	**0.92**	0.03	0.02	0.00	0.01	0.01	0.01	0.00	0.00	0.00
ET	0.04	**0.91**	0.01	0.00	0.00	0.02	0.02	0.00	0.00	0.00
RB	0.01	0.01	**0.91**	0.03	0.03	0.00	0.00	0.00	0.00	0.01
WP	0.00	0.00	0.03	**0.92**	0.03	0.00	0.00	0.00	0.00	0.02
UL	0.01	0.01	0.00	0.02	**0.91**	0.00	0.02	0.02	0.00	0.01
PG	0.01	0.00	0.02	0.00	0.01	**0.92**	0.03	0.01	0.00	0.00
CC	0.00	0.02	0.01	0.00	0.01	0.04	**0.90**	0.00	0.00	0.02
UV	0.00	0.00	0.00	0.00	0.02	0.00	0.00	**0.94**	0.02	0.02
PR	0.00	0.00	0.00	0.01	0.02	0.03	0.00	0.01	**0.91**	0.02
LS	0.00	0.00	0.02	0.03	0.02	0.00	0.00	0.00	0.00	**0.93**
Average classification accuracy rate = **91.7%**										

Note: ∗DR = drink, ET = eat, RB = read book, WP = write on a paper, UL = use laptop, PG = play game, CC = call cellphone, UV = use vacuum cleaner, PR = play guitar, LS = lay on a sofa.

4.3 Precision, Sensitivity, Specificity, and F1-Score

The proposed model has also been evaluated in terms of other evaluation metrics, such as precision, sensitivity, specificity, and F1-score. As given in Eq. (10), the precision of a given interaction class is the number of true positive values out of the total number of positives obtained. Similarly, Eq. (11) shows that sensitivity is the number of true positive values out of the total number of true positives and false negatives obtained. On the other hand, specificity is the number of true negative values out of the total number of true negatives and false positives obtained, as shown in Eq. (12). F1-score is computed using precision and recall values as given in Eq. (13).

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{10}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{11}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{12}$$

$$F1 - score = \frac{2\,(Precision \times Recall)}{(Precision + Recall)} \tag{13}$$

The proposed model has achieved an average precision, sensitivity, specificity, and F1-score of 91.3%, 91.7%, 99.1%, and 91.5% respectively over the MSR Daily Activity 3D dataset respectively as shown below in Table 2.

Table 2: Precision, sensitivity, specificity, and F1-scores over the MSR daily activity 3D dataset

Class	Precision	Sensitivity	Specificity	F1-score
DR	0.91	0.92	0.99	0.91
ET	0.90	0.91	0.99	0.90
RB	0.91	0.91	0.99	0.91
WP	0.92	0.92	0.99	0.92
UL	0.91	0.91	0.99	0.91
PG	0.92	0.92	0.99	0.92
CC	0.89	0.90	0.99	0.89
UV	0.93	0.94	0.99	0.93
PR	0.91	0.91	0.99	0.91
LS	0.93	0.93	0.99	0.93
Mean	0.91	0.92	0.99	0.91

4.4 Body Part Detection Rate

The accuracy of the identification of the twelve body parts that were detected using the graph-based approach is also tested by comparing the obtained coordinates of these points with the ground truth values. The accuracy is given in terms of percentages in Table 3.

Table 3: Body part detection rate achieved over MSR daily activity 3D dataset

Part	DR	ET	RB	WP	UL	PG	CC	UV	PR	LS	AVG
HD	92.23	90.34	90.03	90.12	92.24	93.4	94.32	90.45	94.35	90.12	91.76
RE	95.67	93.03	92.12	90.11	93.56	96.05	92.35	94.32	93.27	96.05	93.65
LE	93.35	95.67	91.78	94.38	96.05	94.38	93.62	92.23	93.56	94.32	93.93
RH	91.45	90.51	91.63	95.67	92.23	94.38	90.56	93.27	95.67	93.27	92.86
LH	97.59	90.12	95.67	91.45	92.35	97.59	92.03	93.56	96.05	92.23	93.86
NK	94.38	96.05	94.38	95.67	93.35	93.35	91.14	92.23	97.59	95.67	94.38
TRS	93.62	92.72	95.67	87.24	94.32	95.67	93.35	95.67	92.23	94.32	93.48
BTR	92.23	95.67	97.59	93.56	94.38	94.38	90.42	92.23	93.56	96.05	94.01
RK	93.35	91.39	94.38	92.23	97.59	92.23	93.24	93.56	94.38	92.23	93.46
LK	93.56	97.59	93.56	95.67	94.32	93.56	90.76	94.32	92.23	94.38	94.00
RF	93.27	91.45	92.23	91.45	94.38	92.23	91.09	94.38	97.59	94.32	93.24
LF	95.67	92.35	95.67	94.38	93.27	93.27	93.56	92.35	94.38	93.27	93.82

Average part detection rate = **93.53%**

Note: ∗HD = head, RE = right elbow, LE = left elbow, RH = right hand, LH = left hand, NK = neck, TRS = upper torso, LTR = bottom torso, RK = right knee, LK = left knee, RF = right foot, LF = left foot, AVG = average.

4.5 Comparison with State-of-the-art Methods

This section compares the proposed model with some other recently developed state-of-the-art (SOTA) models that were tested on the same dataset used in this paper. Table 4 shows the proposed system outperforms them.

Table 4: Comparison with other SOTA methods

Methods	Accuracy (%)
Metric learning autoencoder [35]	67.1
Cross-view action modeling [36]	73.1
Deep moving poselets [37]	84.4
Actionlet ensemble [38]	86.0
Interaction part modeling [39]	89.3
Combined deep architectures [40]	91.3
Proposed method	**91.7**

5 Discussion

This article proposes an efficient HOIR method that uses overall human pose, human body parts, and scene features. The proposed model has achieved impressive results in terms of pose estimation and interaction detection. In this model, all input frames are first pre-processed to enhance their quality. After this, all images are segmented, and human and object pairs are extracted from them. The human silhouette is then used for pose estimation, which involves detecting twelve key human body points. Full images are used to extract scene information, detected human silhouettes are used to extract full body features, and estimated key points are used for the extraction of points-based features. The different kinds of features are then combined and then the feature vector is optimized. Finally, this feature vector is used by HMM to detect accurate human-object interaction.

A quick review of the experimentation and results is as follows: The MSR Daily Activity 3D dataset has been used for experimentation. The proposed model has achieved a mean accuracy of 91.7%. A confusion matrix showing the true positive, true negative, false positive, and false negative values of each class is obtained. Moreover, body part detection rates are also obtained by comparing the detected key points with the ground truth values. An average rate of 93.53% has been achieved which shows that the proposed pose estimation technique is quite efficient. The proposed model also outperforms other state-of-the-art systems tested on the same dataset.

Although the experimental results show that the proposed model is capable of recognizing various human-object interactions accurately, we believe there's still room for improvement. For example, currently, only one feature for each category was extracted. However, more features can be mined for each category to improve the results. This would also mean that the feature vector would get bigger and hence time complexity would increase. So far, we have worked on the RGB videos of the MSR Daily Activity 3D dataset. Since this dataset also includes depth videos, using RGBD videos as input can improve the overall efficiency of the system. One reason for this is that depth information is less affected by lighting conditions as compared to RGB information. The proposed model was tested on one dataset but testing it on more complex datasets can be useful for determining its limitations against various challenges. This dataset is comprised of videos of daily activities. Using different datasets, such

as sports datasets or those with more outdoor scenes will also be useful for establishing the general applicability of the proposed model.

6 Conclusion

In this paper, a human interaction recognition system has been proposed that exploits the human pose and scene features information for accurate classification of different daily activities provided in the MSR Daily activity 3D dataset. The method involved the various steps of preprocessing the input videos, segmenting out the human-object pairs, identifying twelve key human body parts, extracting three different features, optimizing the combined feature vector, and classifying the interactions. Detailed experiments on this model using the above-mentioned dataset have shown that the system is efficient and robust against many challenges. Moreover, it has outperformed some other state-of-the-art solutions that have been tested on the same dataset. Hence, this research proves that using a combination of human poses, human body parts, and scene information is more fruitful as compared to using only one of two of these important elements for human interaction recognition.

Funding Statement: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2018-0-01426) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). In addition; the authors would like to thank the support of the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University. This work has also been supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project Number (PNURSP2022R239), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Also; this work was partially supported by the Taif University Researchers Supporting Project Number (TURSP-2020/115), Taif University, Taif, Saudi Arabia.

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

References

[1] T. A. Shloul, U. Azmat, S. A. Alsuhibany, Y. Y. Ghadi, A. Jalal *et al.,* "Smartphone sensors based physical life-routine for health education," *Intelligent Automation & Soft Computing*, vol. 34, no. 2, pp. 715–732, 2022.

[2] Y. Y. Ghadi, M. Batool, M. Gochoo, S. A. Alsuhibany, T. A. Shloul *et al.,* "Improving the ambient intelligence living using deep learning classifier," *Computers, Materials & Continua*, vol. 73, no. 1, pp. 1037–1053, 2022.

[3] A. Jalal, M. A. K. Quaid and K. Kim, "A wrist worn acceleration based human motion analysis and classification for ambient smart home system," *Journal of Electrical Engineering & Technology*, vol. 14, pp. 1733–1739, 2019.

[4] A. Jalal, A. Nadeem and S. Bobasu, "Human body parts estimation and detection for physical sports movements," in *Proc. C-CODE*, Islamabad, Pakistan, pp. 104–109, 2019.

[5] S. Badar, A. Jalal and M. Batool, "Wearable sensors for activity analysis using SMO-based random forest over smart home and sports datasets," in *Proc. ICACS*, Lahore, Pakistan, pp. 1–6, 2020.

[6] A. Usman, Y. Ghadi, S. Tamara, A. Suliman, A. Jalal *et al.,* "Smartphone sensor-based human locomotion surveillance system using multilayer perceptron," *Applied Sciences*, vol. 12, no. 5, pp. 2550, 2022.

[7] M. J. Hussain, A. Shaoor, S. A. Alsuhibany, Y. Y. Ghadi, T. A. Shloul *et al.,* "Intelligent sign language recognition system for e-learning context," *Computers, Materials & Continua*, vol. 72, no. 3, pp. 5327–5343, 2022.

[8] Y. Y. Ghadi, N. Khalid, S. A. Alsuhibany, T. A. Shloul, A. Jalal *et al.,* "An intelligent healthcare monitoring framework for daily assistant living," *Computers, Materials & Continua*, vol. 72, no. 2, pp. 2597–2615, 2022.
[9] M. Batool, A. Jalal and K. Kim, "Telemonitoring of daily activity using accelerometer and gyroscope in smart home environments," *Journal of Electrical Engineering and Technology*, vol. 15, pp. 2801–2809, 2020.
[10] A. Raza and A. Jalal, "A smart surveillance system for pedestrian tracking and counting using template matching," in *Proc. ICRAI*, Rawalpindi, Pakistan, pp. 1–6, 2021.
[11] S. Tamara, J. Madiha, M. Gochoo, A. Suliman, Y. Ghadi *et al.,* "Student's health exercise recognition tool for e-learning education," *Intelligent Automation & Soft Computing*, vol. 35, no. 1, pp. 149–161, 2022.
[12] M. Waheed, M. Javeed and A. Jalal, "A novel deep learning model for understanding two-person interactions using depth sensors," in *Proc. ICIC*, Lahore, Pakistan, pp. 1–8, 2021.
[13] K. Murphy, A. Torralba and W. Freeman, "Using the forest to see the trees: A graphical model relating features, objects and scenes," in *Proc. NIPS*, Vancouver, Canada, pp. 1–10, 2004.
[14] S. Gupta and J. Malik, "Visual semantic role labeling," arXiv preprint arXiv:1505.04474, pp. 1–11, 2015.
[15] A. Ayesha, Y. Ghadi, M. Alarfaj, A. Jalal, S. Kamal *et al.,* "Human pose estimation and object interaction for sports behaviour," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 1–18, 2022.
[16] K. Nida, 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. 970, 2021.
[17] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. CVPR*, San Francisco, USA, pp. 17–24, 2010.
[18] Y. Y. Ghadi, M. Waheed, T. Al Shloul, S. A. Alsuhibany, A. Jalal *et al.,* "Automated parts-based model for recognizing human–object interactions from aerial imagery with fully convolutional network," *Remote Sensing*, vol. 14, no. 6, pp. 1492, 2022.
[19] N. Khalid, Y. Ghadi, M. Gochoo, A. Jalal and K. Kim, "Semantic recognition of human-object interactions via Gaussian-based elliptical modeling and pixel-level labeling," *IEEE Access*, vol. 9, pp. 111249–111266, 2021.
[20] M. Waheed, A. Jalal, M. Alarfaj, Y. Y. Ghadi, T. Al Shloul *et al.,* "An LSTM-based approach for understanding human interactions using hybrid feature descriptors over depth sensors," *IEEE Access*, vol. 9, pp. 167434–167446, 2021.
[21] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang *et al.,* "Learning human-object interaction detection using interaction points," in *Proc. CVPR*, Seattle, USA, pp. 4116–4125, 2020.
[22] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *Proc. ECCV*, Berlin, Heidelberg, pp. 494–507, 2010.
[23] T. Wang, R. M. Anwer, M. H. Khan, F. S. Khan, Y. Pang *et al.,* "Deep contextual attention for human-object interaction detection," in *Proc. ICCV*, Seoul, Korea, pp. 5694–5702, 2019.
[24] T. He, L. Gao, J. Song and Y. F. Li, "Exploiting scene graphs for human-object interaction detection," in *Proc. ICCV*, Montreal, Canada, pp. 15984–15993, 2021.
[25] A. Gupta, A. Kembhavi and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.
[26] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, pp. 167–181, 2004.
[27] X. Xu, G. Li, G. Xie, J. Ren and X. Xie, "Weakly supervised deep semantic segmentation using CNN and ELM with semantic candidate regions," *Complexity*, vol. 2019, pp. 1–12, 2019.
[28] A. X. Falcao, J. Stolfi and R. de Alencar Lotufo, "The image foresting transform: Theory, algorithms, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 19–29, 2004.
[29] Z. Wang, B. Fan and F. Wu, "Local intensity order pattern for feature description," in *Proc. ICCV*, Barcelona, Spain, pp. 603–610, 2011.

[30] M. A. R. Ahad, M. Ahmed, A. D. Antar, Y. Makihara and Y. Yagi, "Action recognition using kinematics posture feature on 3D skeleton joint locations," *Pattern Recognition Letters*, vol. 145, pp. 216–224, 2021.
[31] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, 2004.
[32] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[33] J. Li, A. Najmi and R. M. Gray, "Image classification by a two-dimensional hidden markov model," *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 517–533, 2000.
[34] J. Wang, Z. Liu, Y. Wu and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. CVPR*, Providence, USA, pp. 1290–1297, 2012.
[35] G. Andresini, A. Appice and D. Malerba, "Autoencoder-based deep metric learning for network intrusion detection," *Information Sciences*, vol. 569, pp. 706–727, 2021.
[36] J. Wang, X. Nie, Y. Xia, Y. Wu and S. C. Zhu, "Cross-view action modeling, learning, and recognition," in *Proc. CVPR*, Columbus, USA, pp. 2649–2656, 2014.
[37] E. Mavroudi, L. Tao and R. Vidal, "Deep moving poselets for video based action recognition," in *Proc. WACV*, Santa Rosa, USA, pp. 111–120, 2017.
[38] J. Wang, Z. Liu, Y. Wu and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 914–927, 2013.
[39] Y. Zhou, B. Ni, R. Hong, M. Wang and Q. Tian, "Interaction part mining: A mid-level approach for fine-grained action recognition," in *Proc. CVPR*, Boston, USA, pp. 3323–3331, 2015.
[40] A. Tomas and K. K. Biswas, "Human activity recognition using combined deep architectures," in *Proc. ICSIP*, Singapore, pp. 41–45, 2017.