Computers, Materials & Continua
DOI: 10.32604/cmc.2022.023553
Article

Tech Science Press

Human Pose Estimation and Object Interaction for Sports Behaviour

Ayesha Arif[1], Yazeed Yasin Ghadi[2], Mohammed Alarfaj[3], Ahmad Jalal[1], Shaharyar Kamal[1] and Dong-Seong Kim[4,*]

[1]Department of Computer Science, Air University, Islamabad, 44000, Pakistan
[2]Department of Computer Science and Software Engineering, Al Ain University, Al Ain, 15551, UAE
[3]Department of Electrical Engineering, College of Engineering, King Faisal University, Al-Ahsa, Saudi Arabia
[4]Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi, Korea
*Corresponding Author: Dong-Seong Kim. Email: dskim@kumoh.ac.kr

Received: 12 September 2021; Accepted: 16 November 2021

Abstract: In the new era of technology, daily human activities are becoming more challenging in terms of monitoring complex scenes and backgrounds. To understand the scenes and activities from human life logs, human-object interaction (HOI) is important in terms of visual relationship detection and human pose estimation. Activities understanding and interaction recognition between human and object along with the pose estimation and interaction modeling have been explained. Some existing algorithms and feature extraction procedures are complicated including accurate detection of rare human postures, occluded regions, and unsatisfactory detection of objects, especially small-sized objects. The existing HOI detection techniques are instance-centric (object-based) where interaction is predicted between all the pairs. Such estimation depends on appearance features and spatial information. Therefore, we propose a novel approach to demonstrate that the appearance features alone are not sufficient to predict the HOI. Furthermore, we detect the human body parts by using the Gaussian Matric Model (GMM) followed by object detection using YOLO. We predict the interaction points which directly classify the interaction and pair them with densely predicted HOI vectors by using the interaction algorithm. The interactions are linked with the human and object to predict the actions. The experiments have been performed on two benchmark HOI datasets demonstrating the proposed approach.

Keywords: Human object interaction; human pose estimation; object detection; sports estimation; sports prediction

1 Introduction

In the digital era, technology is the most significant tool to ease daily human life. Artificial Intelligence (AI) is a vast field of technology used in various research developments of expert systems and computer vision. In automated systems, technology has progressed significantly in the last couple of decades towards the computerization of humans in several applications [1]. Human object interaction is a vast domain and it has many complexities in artificially intelligent systems. Moreover, in a recent study, psychophysicists declared that the understanding of an image or video in a single glimpse is not easy for humans [2]. Although all the social events have been categorized in different fields whereas each field consists of different circumstances. This variety of events requires different kinds of classification between human, object, scene, and background. Therefore, a lot of research has been done in the past and on humans and objects for understanding the events. To detect the human, the authors have considered the human first followed by differentiating it from the background and estimating the pose of the human. After this, they employ object detection and classification techniques.

Event classification and human-object interaction have been used in many applications such as surveillance systems, railways platforms, airports, and seaports where detection of normal and abnormal events along with detection for real-time data is critical [3]. However, there are massive challenges in the way of improving the accuracy of human-object interaction for sports and security agencies that need to identify daily activities [4], office work, gym activities, smart homes and smart hospitals, and understanding of activities in educational institutions. All human activity-based and smart systems need to understand the event and take a decision to arrange activities in a well-organized manner.

In this article, we proposed a unique method for HOI recognition using object detection via Gaussian Matric Model (GMM) and human pose estimation (HPE). We developed a hybrid approach for pre-processing images including salient maps, skin detection, HSV plus RGB detection, and extraction of geometrical features. We designed a system with the combination of Gaussian matrix model (GMM) and k-means to detect the human skeleton, draw ellipsoids of human body parts, and detect the object by a combination of k-means as well as YOLO. A combination of SK-learn and HOG was used for classification and activity recognition. We used two publically available benchmarks datasets for our model and fully validated our results against other state-of-the-art models. The proposed technique processes the data through four different aspects: unwanted components elimination from the image, extraction of hierarchal features, detection of an object based on some factors, and classifiers. In addition, the proposed methodology has been applied to publically available datasets: the PAMI'09 and the UIUC Sports datasets and obtained significant improvement in activities recognition rate over other state-of-the-art techniques.

The rest of the paper is organized as follows. Section 2 contains related work, Section 3 presents the architecture of the proposed model. Section 4 describes the performance evaluation of the proposed work, Section 5 discusses the proposed methodology. Section 6 reports related discussions. Section 6 concludes the paper and provides some future directions.

2 Related Works

In this research article, we discuss human-object interactions using interaction algorithms over sports datasets.

2.1 Human Pose Estimation

Various researches have been done using different approaches to improve the labeling of scenes since it causes false recognition [5]. Designed the object recognition and representation methods that compare the overall pixels to understand the status of the image. Then, they match the kernel to understand the object properly. By combining two different techniques, the MRG technique and the segmentation tree, they show the contextual relationship with respect to detection of the edges by following the connected components [6]. In [7] the adopted an approach is to use depth maps for CRF modeling and system development for scene understanding using less bright images with a simple background. Classification is done on the basis of kernel features [8]. By using depth images, detection and localization of objects in 3D and foreground segmentation of RGB-D images are performed by [9].

2.2 *Action Recognition via Inertial Sensor*

In [10] the extended model of a semantic manifold by combining manually local contextual relationships and semantic relationships to classify the events. Many researchers considered the human object interaction detection by using object detection and human body detection. Human and object attention maps in that approach are constructed using contextual appearance features and local encoding. Only detection of object takes place instead of identication, Use of the complex in the term of time and computationally expensive because of the additional neural network inference [11]. Similarly; hierarchical segmentation proposed by [12] performed contour detection and boundary-detection techniques on RGB images. After the segmentation, a histogram of oriented gradient (HOG) is obtained along with the combination of deformable part model (DPM) for the object detection.

3 Material and Methods

The proposed system based on pre-processes, segmentation and objects detection as initial steps in the input images. After the detection of an object from the image, the human has been detected using salient maps and the pose of the human body has been estimated using GMM [13]. After detecting the human and object, we compute the geometrical features from the segmented images including object centroid, object length, object width, and object acquired area. We take the extreme points of the object (extreme left point, extreme right point, topmost point, and bottom-most point) from the centroid of the object. Next, we apply naive Bayes to find the features from both techniques. On the other hand, from the human body, we detect SK features and full-body features from ellipsoids. Then, to symbolize the reduced features, we optimize the features and find human object co-occurrence. We optimize the features and find co-occurrence between human and object. Finally, the last step is the classification of the event. An overview of the proposed system is shown in Fig. 1.

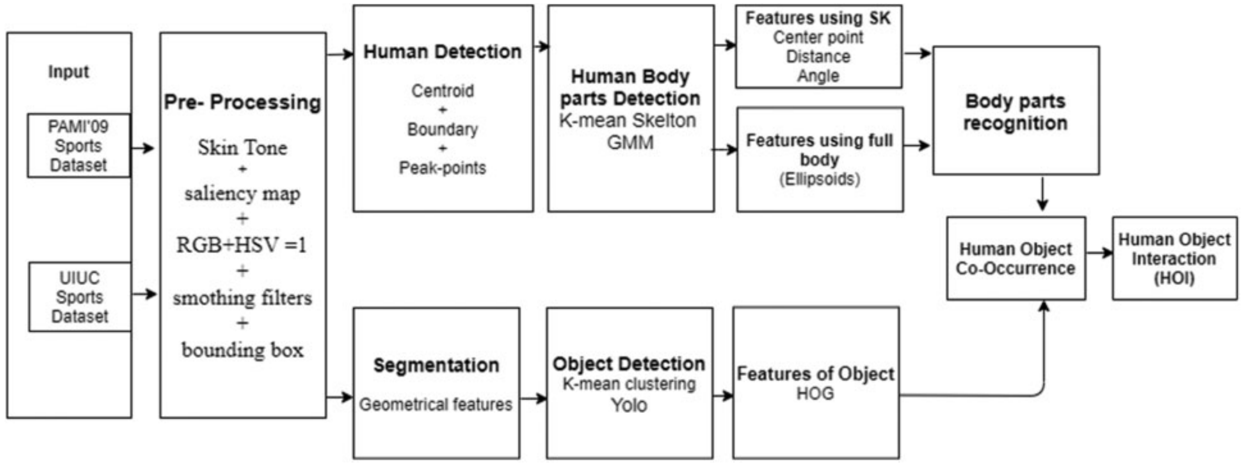

Figure 1: The system architecture of the proposed model of Human Object Interaction (HOI)

3.1 Preprocessing of the Data

First of all, we need segmented noticeable regions to detect the human-object interaction. For this, we perform foreground extraction [14] and smart image resizing. We used the Salient maps method to extract the salient regions and salient object detection as shown in Fig. 2. Low-Rank and LSDM models extract the saliency maps and also capture tree-structure sparsity with the norm. For efficient results, we make partitions of the images and non-overlapping patches. The input image has been divided into N patches $\{Pi\}\ N(i = 1)$.

$$||\mathrm{D}|| = \sqrt{(\mathrm{f}_1\mathrm{x} + \mathrm{f}_2\mathrm{x})^2 + (\mathrm{f}_1\mathrm{y} + \mathrm{f}_2\mathrm{y})^2} \tag{1}$$

we extract the D dimension features for each patch Pi and use a vector $Fi \in R\ D$ for representation. The feature vector $F = \{f_1, f_3, f_3, \ldots f_n\} \in R\ D \times N$ is extracted from the matrix representation of the input image.

$$\left\{ \begin{array}{l} min \\ L, S \end{array} \right. \ ||\mathrm{L}||* \ + \lambda\Omega(\mathrm{S}) \quad \mathrm{s, t} \tag{2}$$

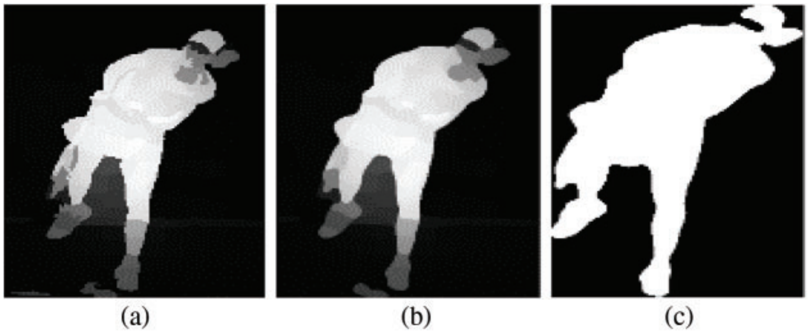

Figure 2: Steps of human silhouette detection. (a) Result of skin tone detection (b) result of salient region detection and (c) result of smoothing filters and bounding box

Therefore, we have designed an algorithm for the decomposition of feature matrix into $(F = L + S)$, some redundant information (L) as well as some structured salient maps **(S)** in which, $\Omega(\mathrm{S})$ is the norm of s-sparsity-inducing. To regularize and preserve the structures, we use the relevant and latent structures and their relationships.

3.1.1 Foreground Extraction

To enhance the silhouette extracted by saliency maps, we perform the segmentation via skin tone detection by using the color-space transformation approach [15] which is achieved by using heuristic thresholds. We extract some skin tone regions by using the $\mathrm{YC_BC_R}$ model. The values of the threshold are specified as R= 0.299, G= 0.287 and B= 0.11.

Through random thresholds of color space transformation, some enhanced regions are extracted. For the chrominance segmentation, we classify the skin and the non-skin regions into different parts precisely.

3.1.2 Smoothing Threshold

We applied filters and some manual thresholds on the foreground to make the resultant image accurate. We have applied a manual threshold to fill the holes and region connector for the small region connection. The threshold range of above 30 is considered as 255 and below 30 as 0.

$$\text{if}(v \geq 30) \tag{3}$$

$$\{ v = -255\} \text{ else } \{ v = 255\}$$

These are the detection steps of human silhouette from images with 2 different techniques and in Fig. 2c both the techniques have been merged.

3.2 Human Detection

The segmentation of the human body leads to the extraction of active regions between human and object. We extract the centroid, obtain the boundary, and highlight the boundary points and eight peak points [16].

3.2.1 Centroid

After the detection of one human body from an image, the spatial feature detection has been performed and the centroid of the human body has been found, which is a torso point Fig. 3a. This helps in detecting the human posture, leading towards finding the action.

3.2.2 Boundary

We performed the boundary extraction algorithms to find the boundary and highlight spatial points on the boundary Fig. 3c. This leads to estimating the human body posture.

3.2.3 Peak Points

After boundary extraction, we find the peak point on the boundary Fig. 3d to estimate the human posture.

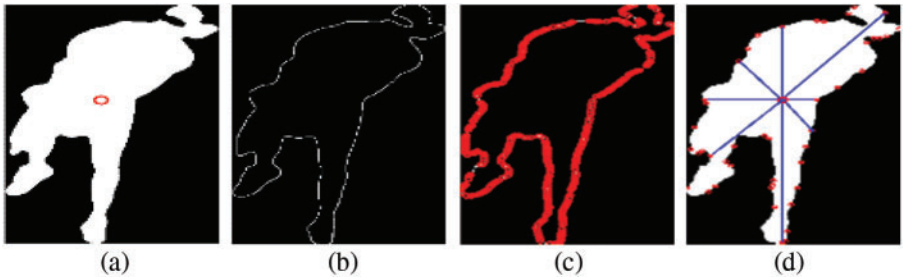

Figure 3: Detection of human body (a) Centroid (b) Boundary (c) Boundary points (d) peak points

3.3 Object Segmentation

The segmentation of objects leads to active regions between humans and objects [17]. We detected the objects using YOLO. Image classification along with localization on each matrix has been performed to predict the bounding boxes and probabilities for objects by extracting geometrical features.

In object segmentation to extract the spatial features, we extract four different extreme points and four features of length, width, centroid, and area of the region. These extracted points and features are in further processing, i.e., the parameter of rectangle or square and area of a rectangle (See Fig. 4). For the measuring of the distance between the x and y extreme points, we use Euclidean distance as

$$||d|| = \sqrt{(Ax - By)^2 + (A'x - B'y)^2} \tag{4}$$

where, $||d||$ represents the distance between two points by using Euclidean, Ax and $A'x$ represent the x-coordinates. While By and $B'y$ represent the y-coordinates.

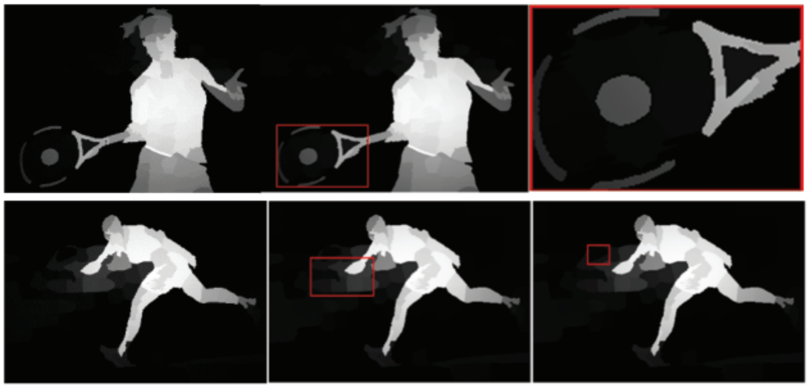

Figure 4: Detection of objects by using K-mean and geometrical features extraction

3.4 Human Body Parts Detection

The skeleton model has been used to detect the human silhouette and by using k-means clustering, circles have been drawn on the human body. GMM model has been used for the ellipse fitting on human body poses by using K number of fixed ellipses for performance prediction. GMM Ellipse fitting algorithm is responsible for drawing the K number of E ellipses for the best coverage of the targeted region $\alpha(E)$.

We use skeleton tree by implementing compact representation for all the skeleton branches, we draw all the possible circles by using the tangent from the central branch. The compact and the lossless representation uses the medial axis transform (MAT) for the denotation of the centroid and the radius which represents the maximum inscribed circle and their coloring. “V” and “W” are the edges of a node of the graph G = (V, W), where V is endpoint node and W is skeleton segment which is the part of Li $\in$ S, S $\in$ {1, . . . , |W|} denoting the number of edges.

$$\mathrm{C} = -\sum_{\mathrm{i=1}}^{|\mathrm{w}|}\sum_{\mathrm{j=1}}^{12} \mathrm{Pij}\log \mathrm{pij} + \log|\mathrm{s}| \tag{5}$$

where Pij is jth bin histogram of the ith node. The term log|S| represents the overall information of the skeleton as shown in Fig. 5.

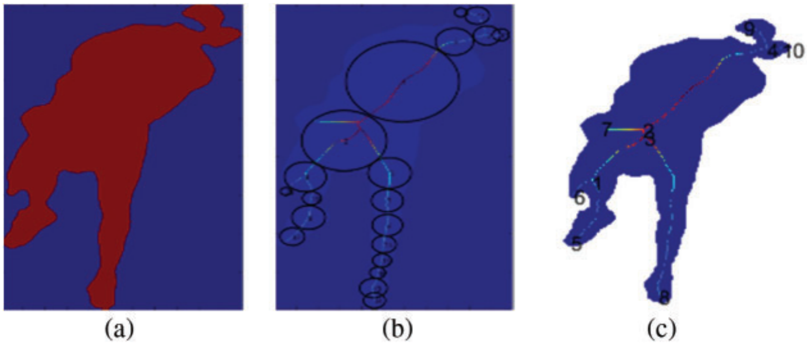

Figure 5: Estimation of skeleton for pose estimation by using skeleton model. (a) 2D image, (b) Medial Axis Transform and (c) Centroid of a circle that is tangent from the boundary

3.5 Object Detection

We used two novel approaches, Fuzzy C-Means and Random forest, for super-pixels and objects segmentation respectively.

3.5.1 Clustering

First of all, clusters are the centroids for each object of the class and then randomly initialized. Feature vectors determine the dimensions of centroids and Euclidian distance is used for assigning the cluster for each object [18]. Additionally, the pixels are assigned to the clusters having the minimum distance from the centroid. These clusters are assigned to classes of objects and the object mean is calculated again to check the difference from the previous value, and the process continues until a constant value is obtained. Fig. 6 represents the clustering. We used 6 classes of PAMI'09 and 8 activities of the UIUC dataset's feature vectors divided into 6 and 8 clusters respectively, which is obtained by some primary steps. To calculate the area Å of a rectangle, we use extreme connected points and connect them to make a rectangle for the object segmentation. For the particular rectangle, the area Å is calculated as:

$$\text{Area of rectangle} = \mathit{Length} \times \mathit{Width} \tag{6}$$

$$\text{Å} = (Q - P) \times (m - n) \tag{7}$$

where Å represents one side of the perimeter, $(Q - P)$ is one side of the rectangle, and $(m - n)$ is the other side of the rectangle. Å symbolizes the area of a rectangle.

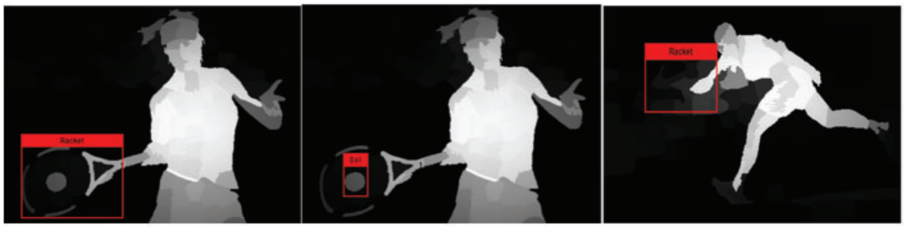

Figure 6: Similar-pixels clustering for object detection

3.5.2 Features Computation Using HOG

We performed segmentation on the images by using edge detection and preserved the edges and used HOG to improve the accuracy.

3.6 Features of Human Pose Using GMM

For the extraction of features, we draw ellipses on the human body using a fixed number of ellipses [19]. GMM is performed for computing parameters and best coverage for ellipses. We did this using a two-step method: drawing the ellipses using k-means and fixing the maximum number K.

3.6.1 Ellipses Using K-Means

The central skeleton is the medial axis as used for the tangent drawing on the boundary and is continuously changing. A 16-bit histogram is used for each circle and the radius of each circle is computed. The circular shape is defined using MAT-based histograms.

3.6.2 Fixing the Maximum Number K

A GMM model is used to draw the ellipsoids on the different body poses by a fixed number K of ellipses E of the image and to predict the performed activity accordingly [20]. GMM-EM algorithm is responsible for computing the parameters for a fixed number k of ellipses E that has achieved the best coverage $\alpha(E)$ in two steps. Ellipses are devolved using the GMM Ellipses fitting model. We wanted to calculate the parameters of ellipses E and by fixing them until K in Algorithm 1 for the best coverage of the region from the silhouette.

Algorithm 1: Ellipse Fitting Algorithm (EFA)

```
Input: EHS: Extracted Human Silhouettes
Input: Binary image I.
Output: Set of ellipses E with the lowest IC
[S,R] = ShapeSkeleton(I)
C = ShapeComplexity(S,R)
CC = InitializeEllipse(S,R)
K = 1
IC* = ∞
Repeat
        SCC = SelectHypothesis(k,CC)
        E = GMM-EM(I,SCC,K)
        IC = ComputeIC(I,E,C)
        ICmin = C. log(1−0.99)+ 2.k
              if IC < AIC* then
              IC* = IC
              E* = E
              End
        K= K+1
Until K==12
```

$P \in F \times G$ is the probability of pixels belonging to the ellipse Ei in the model. Ci is the origin of Ei and Mi is a positive-definite 2×2 matrix which represents the eccentric orientation of Ei. The Gaussian amplitude is used as $Ai = 1$, so the value of probability $Pi(p)$ on the boundary is the same for all ellipses. The possibility that a point belongs to an ellipse Ei is not dependent on the size of the ellipse. We checked this on both datasets as we fixed the number of ellipses by fixing the value of K with the minimum range of 12. This threshold value of K is only applied to get the fixed number of ellipses as represented in Fig. 7.

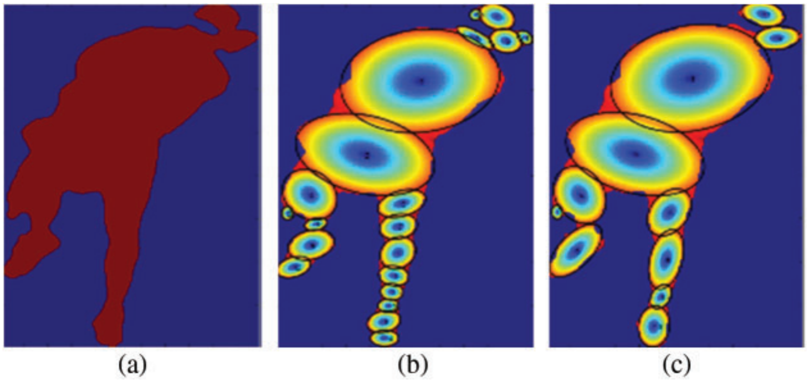

Figure 7: Detection of human body parts by using skeleton model and GMM (a) Ellipsoids from the circles on same centroids by covering all possible pixels (b) Sub-regions of GMM and (c) Limit the number of ellipses k = 12

3.7 Human Object Co-Occurrence

We apply a 3 × 3 convolutional layer to make the vector V as interaction vector of size $\frac{Height}{Size} \times \frac{Width}{Size} \times 2$. At the interface, we extracted the four possible location points for the human body center based on interaction points and interaction vectors.

$$(x_h^n, y_h^n) = (P_x \pm |v_x|,\ P_y \pm |v_y|),\ n = 1, 2, 3\ldots \tag{8}$$

During the training process, the interaction points (IP) and the related human body and the object centroids have fixed geometric structures. While at the inference stage, our generated interaction-points (IP) need to be in grouped-form with the detection of object and their results (bounding-boxes of human-body and object). The points generated by using the interaction points p, center of human h, and center of object o imply a condition on the model: $\boldsymbol{h} \approx \boldsymbol{P} + \boldsymbol{v}$ and $\boldsymbol{O} \approx \boldsymbol{P} - \boldsymbol{v}$.

In Fig. 8, it illustrates the interaction points grouping. It has 3 different inputs including the human body/object bounding boxes (green and red), the interaction points (redpoint) extracted from interaction heat maps, and interaction vectors (IV) in (red arrow) at the location of the interaction points (IP). The four corners of the (green) outlines of interaction boxes (red) are obtained by the given interaction points and the un-signed interaction vectors as shown in Eqs. (9)–(11).

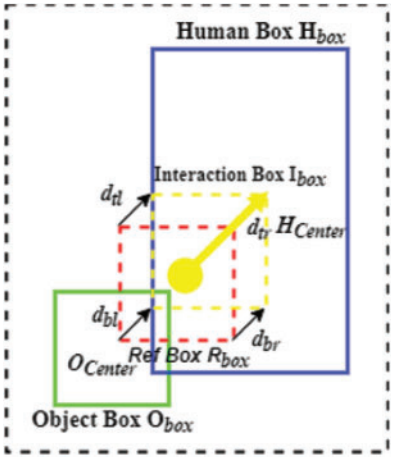

Figure 8: 3 × 3 convolutional layer is used to examine human-object interaction

Fig. 8 represents the procedure of finding the interaction between human and object. The three inputs, namely, the human body/object bounding-box from the object detection branch, the interaction vector is from the interaction points and it is predicted by the interaction vector branch [21]. The current human body/object bounding boxes and their interaction points are regarded as true positive human-object pairs.

$$I(Hbox,\ ibox) > 0 \tag{9}$$

$$I(Obox, ibox) > 0 \tag{10}$$

$$dtl, dtr, dbr, dbl < d\tau \tag{11}$$

Here, *Hbox* and *Obox* are the obtained boxes of humans and objects from the human and object detection. *i* box is the interaction box and it is generated by combining the interaction points and corresponding interaction vectors. *dtl*, *dtr*, *dbl*, and *dbr* are four different vectors with lengths of corners in between the interaction box *ibox* and the reference box *rbox*. dτ is the threshold of vector length set for filtering the negative human-object interaction pairs. The interaction grouping schemes are presented in Algorithm 2.

For the prediction of interaction vectors, we compared point heat maps *P* from the ground-truth with the heat maps ^*P* of all inter-action points. All of these points are with the Gaussian kernel. We used the modified local loss which is proposed in for balancing the positive and negative values. Where Np represents the number of interaction points *IP* in the concerned image. The α and β points are the hyper-parameter points for the contribution of each point. For the interaction points prediction and interaction vector maps V prediction. We use the value of the un-signed vector $V'k = (|Vx|k,\ \ |Vy|k)$ at various interaction points' p and k as the ground-truth. After that, the *L1* loss is used for the related interaction points and here $|Vp|k$ represents the vector and the point which is predicted by loss 1 function.

Algorithm 2: Co-Occurance Human Object Interaction(HOI)

```
Input: Human/Object detector H,O.
   Interaction points and vectors P,V
   Human, Object and Interaction threshold H$\tau$, O$\tau$, I$\tau$
Output: Final Human Object Interaction box If
//Interaction point p makes point set P
//Interaction vector v makes set of vectors V
for Hbox ∈H$\tau$, Oscore ∈O$\tau$, p ∈P
do
    if H scorebox >H $\tau$, O score box > O$\tau$, p score > p$\tau$
    then
    //Interaction box ibox
   //calculate reference box rbox and Obox
       if Hbox, Obox, ibox, rbox satisfy condition 2
       then Sf ← Hscore . Oscore . pscore
       // output the current HOI
      Score Sf
      End if
   End if
End if
```

Meanwhile, we used the point heat map P and ground heat map P' for the prediction and all of them are defined by the Gaussian kernel.

$$L_p = \frac{-1}{Np}\begin{cases} (1 - P_{lmn})^{\alpha} \log P_{lmn}, & if\ P'_{lmn} = 1 \\ (1 - P'_{lmn})^{\beta} (p_{lmn})^{\alpha} \log 1 - P_{lmn}, & \end{cases} \tag{12}$$

In (Eq. (14)), Np represents the number of interaction points and $\alpha\ and\ \beta$ are the parameters that control the contribution of every point. For interaction, we use the vector maps V by using the value of the interaction vector at the interaction point IP as the ground truth. The interaction vector V is $v_i = (|v_x|i, |v_y|i)$ and the loss $L1$ is directed for all the corresponding interaction points.

$$L_m = \frac{1}{N}\sum_{i=1}^{N} |V_{pi} - v'_i| \tag{13}$$

where V_{pi} represents the interaction vectors V at the IP point. The loss function is:

$$L_{total} = L_p + \partial_m L_m \tag{14}$$

Here, ∂_m is a weight for all vector loss terms. Here, we simply specify $\partial_m = 0.1$ for all the experiments.

4 Experiments and Results

This section is organized into five sub-sections. First, two benchmark datasets are described in detail. Second, results evaluation is discussed. Third, human pose estimation is discussed. Fourth, estimation of human-object interaction is explained, and fifth, our proposed work is compared with other state-of-the-art advanced deep learning techniques.

4.1 Datasets Description

To evaluate the performance of the proposed system, we used images based benchmark datasets, namely, the PAMI’09 sports dataset and the UIUC sports dataset contains vast range of backgrounds and verity of sports. These datasets are further divided into testing sets for experiments and testing purposes. These two datasets have been used to detect human bodies and objects and to find the interactions between humans and objects. Both datasets are further classified into different classes of different sports and activities to recognize the different outdoor and indoor activities.

4.1.1 PAMI’09 Dataset

This dataset contains six classes and each class has thirty train, thirty ground truth, and twenty test images. The PAMI2009 dataset contains 480 images with few annotations [22]. Each class has 80 images including 30 images for training, 30 for ground truth, and 20 for testing. Each picture has been cataloged with 12 ellipsoids.

4.1.2 UIUC Dataset

The UIUC sports dataset consists of eight sports activities. In each class, there are 100–240 images. This dataset is comprised of 2000 images, mainly of sportsmen and sportswomen [23].

4.2 Results Evaluations

For efficient results, the dataset has been provided to the Gaussian mixture models in batches of classes. To minimize reconstruction errors, we set the number of training samples according to cross-validation.

4.2.1 Experiment I: Human Pose Estimation

The accuracy of human pose estimation has been measured using Euclidean distance from ground truth [22] of the dataset, which is explained in Eq. (15).

$$\boldsymbol{D}x = \sqrt{\sum_{n=1}^{N} \left(\frac{\boldsymbol{ln}}{\boldsymbol{sn}} - \frac{\boldsymbol{ln}}{\boldsymbol{sn}} \right)^{2}} \tag{15}$$

where the ground truth of dataset *X* is the position of human body parts. *D’* is the threshold, which is 12, and it is used to measure the accuracy between the ground truth and our model.

$$\boldsymbol{D}' = \frac{100}{\boldsymbol{n}} \left[\sum_{n=1}^{K} \left\{ {1\ if\ D \le 15 \atop 0\ if\ D \le 15} \right\} \right] \tag{16}$$

In Tab. 1, columns 2 and 4 represent the distances from the dataset’s ground truth whereas columns 3 and 5 show the human body part recognition accuracies over the PAMI’09 and UIUC sports datasets respectively.

Tabs. 2 and 3 represent the mean accuracy of both the datasets respectively.

Table 1: Human body key point's detection accuracy

Body portions	PAMI'09 dataset		UIUC dataset	
	Distance	Accuracy (%)	Distance	Accuracy (%)
Head	10.8	91	11.32	94
Right upper arm	9.81	82.3	10.81	88
Right lower arm	10.1	86	12.87	89
Left upper arm	10.6	88	9.98	83
Left lower arm	9.93	84	7.45	73
Torso	14.41	98	14.23	96
Right upper leg	13.71	93	9.15	78
Right lower leg	12.78	89	11.89	81
Left upper leg	9.63	78	8.63	75
Left lower leg	11.87	81	10.89	88
Right foot	12.65	90	13.93	92.4
Left foot	13.88	92	14.21	95.7
Mean accuracy rate		87.6		86.09

Table 2: Mean recognition accuracy of PAMI'09 sports dataset

Sports	Cricket batting	Cricket bowling	Volley ball	Tennis serve	Tennis forehand	croquet
Accuracy	74.7	89.2	83.2	91.6	91.7	84.3
Mean = 85.7%						

Table 3: Mean recognition accuracy of UIUC sports dataset

Sports	Badminton	Polo	Croquet	Bocce	Snow boarding	Sailing	Rock Climbing
Accuracy	74.7	89.2	83.2	91.6	91.7	84.3	84.3
Mean = 85%							

4.2.2 Experiment II: HOI

For human-object interaction (HOI) detection and prediction, we use Hourglass as a feature extraction method for pre-training. We randomly initialized the network for generating the interaction points and vectors. During the training of the system, we resize the input images to a resolution of 512 × 512. Standard data augmentation techniques have been employed and an Adam optimizer has been used for the optimization of the loss function during training. Through the testing phase, we perform the flip augmentation method to get final detections and predictions. Moreover, we use a batch with the size of 30 and a learning rate of 2.5.

For the detection branch, we go after the previously proposed HOI estimation methods and employ the Faster R-CNN method with the ResNet-50-FPN and pre-train it on the UIUC training dataset. To acquire the bounding-boxes at the inference, we have set the score thresh-hold for the human to be greater than 0.4 and for the object, it is 0.1. When the interaction box is generated by our interaction points and vectors. The generation of interaction systems has taken about 7s. The interaction group we have has the complexity of $O(Nh\ No\ Ni\ )$, where $Nh,\ \ No,\ \ Ni$ is the number of humans, objects, and interaction points, respectively. In the testing, our grouping scheme is time-efficient and takes less than 2s ($<$ 20% of total time).

4.2.3 Experiment III: Classification of HOI

By following the standard evaluation and testing methods as performed in [24] to analyze our proposed approach, the results are assembled in the form of role-mean-average precision (mAProle). In role-mean-average precision (mAProle), we apply the HOI model and perform it in a way that if and only if one HOI triplet is rewarded as a true-positive when both of the bounding boxes have IoU intersection-over-union (union of interactions) greater than or equal to 0.5 with the labeled data (ground-truth) [25] and the linked interaction class is accurately classified. Firstly, we compare our proposed technique with other state-of-the-art techniques in the literature. Tab. 4 represents the comparison on the PAMI'09 and UIUC dataset. The existing approaches utilize human and object features in multi-stream architecture.

Table 4: State-of-the-art comparison (in terms of mAProle) on the PAMI'09 and UIUC datasets, our approach by combining the HOI and IG with the mAProle of 53.6

lMethods	mAP$_{\text{role}}$
InteractNet	40.0
HOI	45.9
DCA	47.3
BAR	41.1
PMFnet	52.0
IG	52.3
Ours= HOI+IG	53.6

The work of denoted in Tab. 4 as DCA, introduces an interactive network to put on non-interaction suppression and reports with a mAProle of 48.3. Our technique achieves state-of-the-art performance by comparing it to existing techniques with a mAProle of 53.4. Fig. 9 shows that our results are improved by comparing them (mAProle of 53.6) in first pre-training our model on PAMI'09 and UIUC datasets and then fine-tuning and pre-training the model on both datasets.

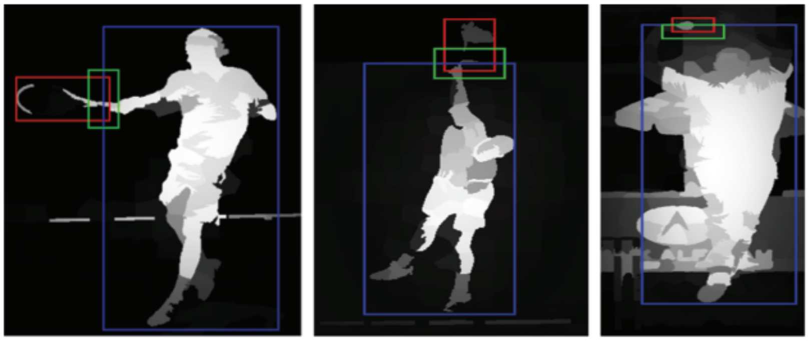

Figure 9: Some results by using Human Object Interaction (HOI) model

4.2.4 *Experiment IV: Qualitative Analysis of our Proposed System*

Finally, after the classification and recognition of human lifelong activities those are performed in this phase. Tab. 5 shows the accuracy of different classes in the form of confusion matrix of PAMI'09 dataset with 90.0% of mean accuracy. This shows the significant improvement and better results from the proposed methodology.

Table 5: Confusion matrix table on PAMI'09 sports dataset

	Cricket batting	Cricket bowling	Volleyball	Tennis serve	Tennis forehand	Croquet
Cricket batting	**9.0**	0	0	0	0	1.0
Cricket bowling	0	**9.6**	0	0.4	0	0
Volley ball	0	0.2	**9.6**	0.2	0	0
Tennis serve	0	0.8	0	**9.2**	0	0
Tennis forehand	0.4	0	0	0	**8.6**	1.0
Croquet	0.2	1.0	0	0.6	0.2	**8.0**
Mean accuracy = 90.0%						

After that, the classification and recognition of human activities are performed over the UIUC sports dataset set. Tab. 6 shows the accuracy of different classes in the form of confusion matrix of the UIUC sports dataset with 87.71% of mean accuracy, which shows significant improvement and better results from the proposed methodology.

Table 6: Confusion matrix table on UIUC sports dataset

Badminton	Polo	Croquet	Bocce	Snow boarding	Sailing	Rock climbing	Badminton
Badminton	**9.0**	0	0	0	0	0	1.0
Polo	0	**8.8**	0	0.8	1.0	0	0
Croquet	0	0.2	**9.6**	0.2	0	0	0
Bocce	0	0.8	0	**8.2**	0	1.0	0
Snow boarding	0	0.4	0	0	**9.0**	0	0.6
Sailing	0	0	0	0.6	0.4	**9.0**	0
Rock climbing	0	0	2.0	0	0.2	0	**7.8**
Mean accuracy = 87.71%							

5 Conclusion

We proposed a novel approach to estimate the HOI in images. Our approach refers to HOI estimation as a fundamental problem of research work in which we perform the pose estimation using skeleton model and GMM. After that, we detect the object by combining the features of K-means clustering and YOLO. Moreover, we generate the interaction points and interaction vectors by using key-point detection and pair those interaction points and vectors with the human and object by using the bounding boxes. HOI interaction was performed by using the HOI interaction group method. Through reference boxes and reference vectors, we estimate the interaction. Our experiments are performed on two HOI benchmark sports datasets, PAMI'09 and UIUC. Our approach outperforms state-of-the-art methods on both datasets with the accuracies of 90.0% and 87.71%, respectively.

In the future, we will extend the interaction vector concept by using multiple vectors from the interaction point to the human body and object to improve the results of our model. We also aim to implement this model in other applications and indoor HOI datasets.

Funding Statement: This research work was supported by Priority Research Centers Program through NRF funded by MEST (2018R1A6A1A03024003) and the Grand Information Technology Research Center support program IITP-2020-2020-0-01612 supervised by the IITP by MSIT, Korea.

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

References

[1] A. Jalal, Y. Kim, S. Kamal, A. Farooq and D. Kim, "Human daily activity recognition with joints plus body features representation using kinect sensor," in *Proc. Int. Conf. on Informatics*, Electronics & Vision, Fukuoka, Japan, pp. 1–6, 2015.

[2] S. A. Rizwan, A. Jalal and K. Kim, "An accurate facial expression detector using multi-landmarks selection and local transform features," in *Proc. Int. Conf. on Advancements in Computational Sciences (ICACS), 2020*, Lahore, Pakistan, pp. 1–6, 2020.

[3] A. Jalal, A. Nadeem, and S. Bobasu, "Human body parts estimation and detection for physical sports movements." in *IEEE 2nd Int. Conf. on Communication, Computing and Digital Systems (C-CODE)*, Islamabad, Pakistan, pp. 104–109, 2019.

[4] N. Khalid, M. Gochoo, A. Jalal, & K. Kim, "Modeling two-person segmentation and locomotion for stereoscopic action identification a sustainable video surveillance system," *Sustainability*, vol. 1, no. 2, pp. 970, 2021.

[5] X. Song, S. Jiang and L. Herranz, "Joint multi-feature spatial context for scene recognition on the semantic manifold," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, Beijing, China, pp. 1312–1320, 2015.

[6] L. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Proc. Int. Conf.on Computer Vision*, Rio de Janeiro, Brazil, pp. 1–8, 2007.

[7] A. Jalal, S. Kamal and D. S. Kim, "Detecting complex 3D human motions with body model low-rank representation for real-time smart activity monitoring system," *KSII Transactions on Internet and Information Systems*, vol. 12, no. 3, pp. 1189–1204, 2018.

[8] S. M. Abid Hasan and K. Ko, "Depth edge detection by image-based smoothing and morphological operations," *Journal of Computational Design and Engineering*, vol. 3, no. 3, pp. 191–197, 2016.

[9] A. Jalal, I. Akhtar and K. Kim, "Human posture estimation and sustainable events classification via pseudo-2D stick model and k-ary tree hashing," *Sustainability*, vol. 12, no. 23, pp. 9814, 2020.

[10] R. Kirti and A. Bhatnagar, "Image segmentation using canny edge detection technique," *International Journal of Techno-Management Research*, vol. 4, no. 4, pp. 8–14, 2017.

[11] S. Gupta, P. Arbeláez, R. Girshick and J. Malik, "Indoor scene understanding with rgb-d images: bottom-up segmentation, object detection and semantic segmentation," *International Journal of Computer Vision*, vol. 112, no. 2, pp. 133–149, 2015.

[12] A. Jalal, M. Batool and S. B. ud din Tahir, "Markerless sensors for physical health monitoring system using ECG and GMM feature extraction," in *Int. Bhurban Conf. on Applied Sciences and Technologies (IBCAST), 2021*, Islamabad, Pakistan, pp. 340–345, 2021.

[13] M. Gochoo, I. Akhter, A. Jalal and K. Kim, "Stochastic remote sensing event classification over adaptive posture estimation via multifused data and deep belief network," *Remote Sensing*, vol. 13, no. 5, pp. 912, 2021.

[14] B. Martinkauppi, M. Soriano and M. Pietikainen, "Detection of skin color under changing illumination: a comparative study," in *Proc. Int. Conf. on Image Analysis and Processing*, Mantova, Italy, pp. 652–657, 2003.

[15] M. Shehnaz and N. Naveen, "An object recognition algorithm with structure-guided saliency detection and SVM classifier," in *Proc. Int. Conf. on Power, Instrumentation, Control and Computing*, Thrissur, India, pp. 1–4, 2015.

[16] A. Jalal, A. Nadeem and S. Bobasu, "Human body parts estimation and detection for physical sports movements," in *Int. Conf. on Communication, Computing and Digital systems (C-CODE), 2019*, Islamabad, Pakistan, pp. 104–109, 2019.

[17] J. -F. Hu, W. S. Zheng, J. Lai, S. Gong and T. Xiang, "Recognising human-object interaction via exemplar based modelling," in *Proc. Int. Conf. on Computer Vision*, Sydney, Australia, pp. 3144–3151, 2013.

[18] 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, no. 4, pp. 1733–1739, 2019.

[19] A. Arif and A. Jalal, "Automated body parts estimation and detection using salient maps and Gaussian matrix model," in *Proc. Conf. on Applied Sciences and Technologies (IBCAST), 2021*, Islamabad, Pakistan, pp. 667–672, 2021.

[20] S. Thual, A. J. Majda and S. N. Stechmann, "A stochastic skeleton model for the MJO," *Journal of the Atmospheric Sciences*, vol. 71, no. 2, pp. 697–715, 2014.

[21] D. K. Vishwakarma, "A Two-fold transformation model for human action recognition using decisive pose," *Cognitive Systems Research*, vol. 61, pp. 1–13, 2020.

[22] K. Buys, C. Cagniart, A. Baksheev, T. De Laet, J. De Schutter *et al.*, "An adaptable system for RGB-D based human body detection and pose estimation," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 39–52, 2014.

[23] M. Pervaiz, A. Jalal and K. Kim, "Hybrid algorithm for multi people counting and tracking for smart surveillance," in *Proc. Conf. on Applied Sciences and Technologies (IBCAST), 2021*, Islamabad, Pakistan, pp. 530–535, 2021.

[24] M. Javeed, A. Jalal and K. Kim, "Wearable sensors based exertion recognition using statistical features and random forest for physical healthcare monitoring," in *Proc. Conf. on Applied Sciences and Technologies (IBCAST), 2021*, Islamabad, Pakistan, pp. 512–517, 2021.

[25] I. Akhter, A. Jalal and K. Kim, "Pose estimation and detection for event recognition using sense-aware features and adaboost classifier," in *Proc. Int. Bhurban Conf. on Applied Sciences and Technologies*, Islamabad, Pakistan, pp. 500–505, 2021.