IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 July 2023, accepted 6 September 2023, date of publication 11 September 2023, date of current version 20 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3314341

RESEARCH ARTICLE

Human-Based Interaction Analysis via Automated Key Point Detection and Neural Network Model

ISRAR AKHTER[1], NAIF AL MUDAWI[2], BAYAN IBRAHIMM ALABDULLAH[3], MOHAMMED ALONAZI[4], AND JEONGMIN PARK[5]

[1]Department of Computer Science, Bahria University, Islamabad 44000, Pakistan
[2]Department of Computer Science, College of Computer Science and Information System, Najran University, Najran 55461, Saudi Arabia
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, 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 Computer Engineering, Tech University of Korea, Siheung-si, Gyeonggi-do 15073, South Korea

Corresponding author: Jeongmin Park (jmpark@tukorea.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation (NRF) (2021R1F1A1063634) funded by the Ministry of Science and Information & Communications Technology (MSIT), Republic of Korea. Also, Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R440), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. In addition, 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/12/40).

ABSTRACT The human interaction with an object is one of the most challenging domains in real-life applications, such as smart homes, surveillance, medical, education, safety-based application of computer vision, and artificial intelligence. In this research article, we have proposed a framework for human and object interaction in real-life examples such as sports and other activities. Initially, we reviewed video-based data by considering the three state-of-the-art data sets. Preprocessing steps have been followed to avoid extra costs, such as video-to-frame conversion, noise reduction and background subtraction. Human silhouette extraction has been performed via the Gaussian mixture model (GMM) and supper pixel model. Next, human body points and object location detection were performed. Finally, human and object-based features have been extracted. To minimize the features replication and to achieve optimized results, we have applied stochastic gradient descent and Restricted Boltzmann Machine; As a result, we have achieved an accuracy of 88.46%, 82.00%, and 88.30% on human body parts recognition over the MPII dataset, UCF_aerial dataset, and wild Dataset respectively. The classification accuracy for the MPII dataset is 92.71%, for the UCF_aerial dataset is 90.60%, and for sports video in the wild Dataset is 92.42%. We have achieved a high accuracy rate compared to other state-of-the-art methods and frameworks due to the complex feature extraction and optimization approach.

INDEX TERMS Features optimization, human-object interaction analysis, human body key points detection, restricted Boltzmann machine, stochastic gradient decent, skeletonization, trajectories.

I. INTRODUCTION

In recent decades, Human-Object Interaction (HOI) identification has attracted many purposes to inform as it plays an essential role in living person scene interpretation. References [1], [2], [3], [4], and [5]. Human-object interaction is crucial in various fields, such as smart homes, surveillance, security, medical applications, sports, video surveillance, intelligent thermostats, e-learning, individual counting, person mapping, and motion detection. In smart homes, human-object interaction facilitates the control and automation of multiple devices, thereby augmenting convenience and energy efficiency. It enables real-time surveillance, face recognition, and access control in surveillance and security systems. It facilitates remote patient monitoring and personalized healthcare in medical applications. It facilitates motion tracking and analysis for performance improvement in sports. It helps identify and monitor individuals through video surveillance [6], [7], [8], [9]. Intelligent thermostats optimize energy consumption

The associate editor coordinating the review of this manuscript and approving it for publication was Christos Anagnostopoulos.

based on occupancy by leveraging human-object interaction. E-learning platforms utilize it for interactive learning experiences. Individual counting, person mapping, and motion detection are improved for accurate monitoring and analysis by human-object interaction. Overall, human-object interaction is a crucial aspect of these domains, as it improves functionality, efficiency, and security [10], [11], [12]. Therefore, human activity recognition (HAR), human-human interaction (HHI), and human-object interaction (HOI) is recognized as demanded subjects in the domains of video processing and intelligent systems (IS). Nevertheless, the scope of this investigation is mostly confined to the HOI classification. Although academics have built numerous effective HOI detection systems in recent years, the work remains challenging, and there is still space for advancement [13], [14], [15], [16], [17]. In addition, the combination of visual, wireless, and depth technologies has facilitated the ability of current HOI identification systems to increase productivity and incidence, especially to deal with experiencing substantial [18], [19], [20].

In increasing public and confidential settings, it is crucial to maintain and recognize social interactions [21], [22], [23]. Human interconnections have two forms: human-human interaction (HHI) and human-object interaction (HOI). HOI includes the actions conducted by an individual in connection to an artifact, whereas [24], [25], [26] HHI pertains to the activities undertaken between two individuals. Recognizing sophisticated human- object connections is crucial in most observations, monitoring systems, supported dwelling, retraining, and e-learning platforms.

Large-scale, demanding, and accessible to the public HOI collections have been produced due to the increased interest of academics in this subject. Scholars have made significant progress and performance in several existing HOI monitoring systems exhibit promising performance [27], [28], [29] Chalearn LAP encounters on self-reported behavior recognition and non-verbal actions prediction during dyadic social contacts: [30], [31], [32], [33] Dataset, strategy, and values. However, these systems are less effective in real-world circumstances due to various differences in scale, personality, brightness disparity, crowded surroundings, and varying perspectives. The significant issues of remote detecting images are; rapid camera mobility, low higher resolution, and compact size of objectives [34].

Most contemporary HOI detection methods employ a multi-stream framework for interaction classification. Typically, the multi-stream framework comprises three distinct streams: human, object, and paired. The human and object sources encapsulate the appearance characteristics of humans and objects, respectively, while the pairwise stream encodes the spatial connection between humans and objects. Individualized scores from the three inputs are then combined for relationship recognition using a late fusion technique [35].

This paper presents a technique for detecting HOIs in RGB and remote sensing imagery. Initially, we have taken RGB data as the effort for the proposed HOI structure. The preprocessing has been applied to reduce the computational rate and resource management. Finally, motion blur noise reduction, frame conversion, resizing and RGB image to binary conversion have been used on the resultant image. The next step is to detect a human from the given images. We applied two methods to see a human in RGB images, namely the supper pixel and GMM model; after that human shape-based model was used to extract complete information about the human silhouette. Human body key points and object detection is the next step and after the recognition of human body key points and objects, we applied features and trajectory extraction approaches. To reduce data and optimize data, we applied stochastic gradient descent, and for classification, we applied Restricted Boltzmann Machine. The following are the key achievements of this study paper:

- We have presented an effective and optimized way of human silhouette extraction process in RGB data using supper pixel, GMM, and shape-based model.
- A robust way to detect human body key themes and object detection has been applied with the support of these points.
- Stochastic gradient descent has been applied to deal with extra information and computational cost optimization.

The remaining segments are arranged as follows: Section II describes and evaluates the literature related to the project system. Section III outlines the system's general technique, which includes a comprehensive pre-classification procedure. Section IV discusses the Dataset utilized in the proposed method and demonstrates the system's stability through several trials. Section V summarizes the data and discusses future studies. Fig. 1 shows the complete route map of the proposed method.

II. RELATED WORK

In recent decades, scientists have been planning to develop HOI recognition technology. They have also examined the behaviors of their methods using distance photos, depth movies, and RGB+D (red, green, blue, and depth) films, in addition to the color image. In most instances, the complete image has been used as a source, and its characteristics are retrieved. This method is straightforward and highly beneficial to identifying many HOI connections in a single picture. Hence, one can extract people and element pairs from the photos and harvest their attributes independently. Furthermore, the additional phases of posture assessment and human body component identification for enhanced image retrieval are typical. This method gives more accurate findings because of the item's added environmental data and human traits.

Consequently, associated research can be divided into appearance and specific examples of HOI person identification. InteractNet, which enhances the Faster R-CNN paradigm with an additional branch to acquire the interaction-specific probability map over target places, is introduced in [36]. Qi et al. [37] present a graph convolution neural network approach and view the HOI challenge as a graph-based

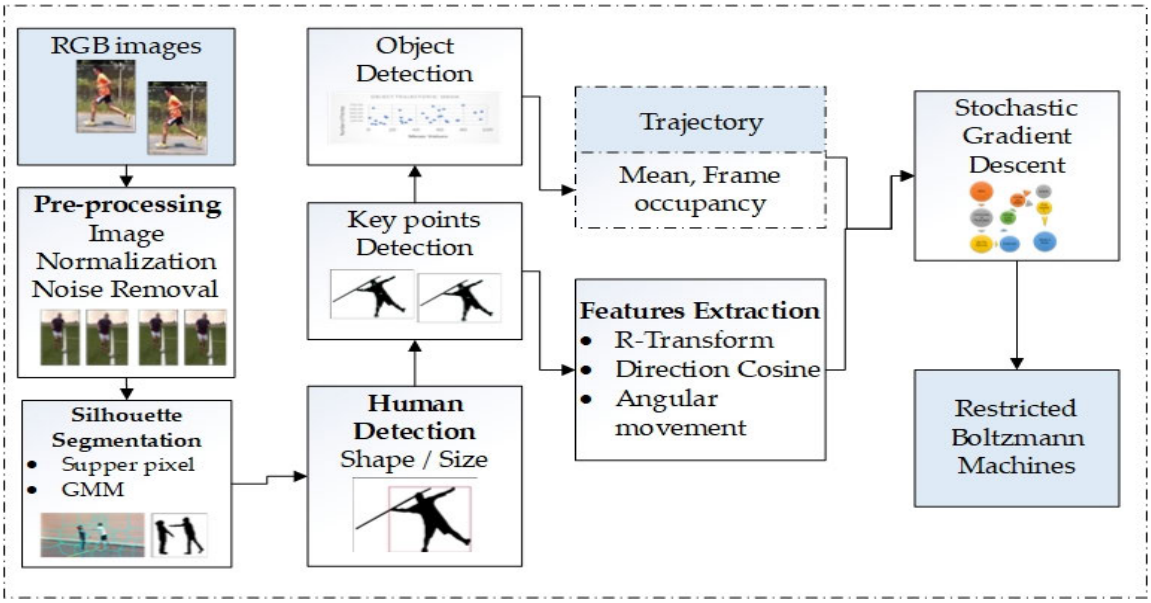


FIGURE 1. Detailed overview of the proposed architecture via data optimization and restricted Boltzmann machines.

optimization issue. Chao et al. The surrounding frames calculated by an already trained detector (e.g., FPN [38]) and the initial image are the inputs to this multi-stream framework. In such a multi-stream building design, the human and entity streaming are based on appearing attributes taken from the backbone network in order and produce estimates with significant trust on the recognized human and object boundary boxes.

A. HUMAN AND HUMAN INTERACTION

This article presents several newly designed HIR systems. Khan et al. [39] suggested a dynamic part-based modeling approach to recognize and identify the body components of a person in adjacent frames. Their method subsequently analyzed newborns' movements to discover a variety of behavioral abnormalities. Researchers gathered the data using Microsoft Kinect in a nearby hospital, although it was simply RGB data. Khan et al. [40] presented a technique for analyzing the body motions of a patient undergoing Vojta treatment. They suggested using color characteristics and pixel positions to partition the patient's body into photos. Then, they classified the right motions by employing SVM and a multifunctional feature map. Jalal et al. [41] also employed a combination of four candidate regions: Spatio-temporal characteristics, energy-based functions, structural rotational and architectural features, and Movements Orthogonal Histograms of Oriented Gradients (MO-HOG). Furthermore, such methods exclusively utilized 2D parameters.

RCNN is an adaptation of a neural network convolution (RCNN) that uses more than one location for segmentation. Scientists attained great precision by collaboratively developing the initiative classifiers and the processing elements. For example, Shen et al. [42] has accurately defined the link between a phrase and an objective using zero-shot acquisition. Lacking contextual information, their system was reviewed using a simple derivational relationship. Furthermore, these approaches has been characterized by intricate characteristics, and their systems possess another high-time sophistication. Numerous recent studies have attempted the instantiation technique to recognize interpersonal interaction.

B. HUMAN OBJECT INTERACTION

After refining the raw photos, the authors have segregated the persons and complex functionalities in the given exchanges. The humanoid creature has been obtained using the Euclidean distance transforms, and individual components have subsequently been retrieved from the vertebrae. The authors also produced irregular facial animations using the GMM (Gaussian mixture model) and labeled the images of human physical sections using the CRF (conditional random field) template. In addition, Yan et al. [43] introduced an HOI recognition system utilizing a neural network with several tasks. Finally, researchers provided a computerized glove named "WiseGlove" to detect hand movements. The system utilized YOLO (you only look once) for feature extraction and a deep convolutional network for connection recognition. The researchers used RGB and skeletal information in their experiments to get a high classification rate. Nevertheless, the data set contains just eight movement classifications. In addition, their technology could only work with a few predetermined components.

Wang et al. [44] suggested using the potential to affect interpersonal interaction significantly. They posed HOI as a question of crucial point determination. To analyze interpersonal interaction, scientists employed a deeply connected strategy. Using the expected engagement locations, the engagement was localized and classified. Gkioxari et al. [36]

discovered individual, verb, and object triplets by focusing on humans based on their physical appearance and execution compactness. Researchers utilized two RGB datasets to validate their technique. Likewise, Li et al. [45] introduced a 3D posture estimation technique and a new standard called ''Ambiguous-HOI.'' To mine characteristics, they utilized 2D and 3D visualization structures. In addition, humans and objects were represented using a cross-modal reliability task and a collaborative learning model. To demonstrate the efficacy of their method, they conducted exhaustive tests on different datasets. Ahmad et al. [46] used an instantiation method to create a gait event categorization system that identified activities using a variety of parameters, such as regular and no periodic movements, motion orientation and movement, horizontal rotational movements, and degrees of freedom.

The topic of the proposed research paper is the detection of HOIs in RGB and remote sensing imagery. The article describes a technique that optimizes computational efficiency by combining preprocessing stages such as motion blur noise reduction, frame conversion, resizing, and RGB to binary conversion. Human detection uses super pixel, Gaussian Mixture Model (GMM) techniques, and shape-based models to extract comprehensive information about human silhouettes. Utilizing critical points of the human body and object detection, features and trajectory extraction techniques are then employed. Stochastic gradient descent is used for data reduction and optimization, whereas Restricted Boltzmann Machine is employed for classification. Using super pixelc, GMM, and shape-based models, the proposed work distinguishes itself by providing an efficient and optimized method for human silhouette extraction from RGB data. Detecting human body key points and subsequent object detection enhances the system's capabilities. Applying stochastic gradient descent facilitates the processing of additional data and the reduction of computational costs. This research demonstrates advancements in HOI recognition by integrating multiple techniques and obtaining promising results in detecting and analyzing human-object interactions.

III. SYSTEM METHODOLOGY

In this subpart, we have discussed the complete procedure of the system method with detailed results, algorithms, and equations. Initially, we took RGB video data as input, the preprocessing step was used to reduce extra cost and motion blur noise, human silhouette extraction was performed, and human and object detection was achieved by various algorithms, after that we have extracted features and trajectories for data optimization, we applied stochastic gradient decent and for classification and activity analysis we applied restricted Boltzmann machine.

A. DATA PREPROCESSING

Prior human authentication and identification, we applied numerous preprocessing strategies for reducing computationally cost and duration. This involves the experimental

(a) (b)

FIGURE 2. Example images (a) original RGB image and (b) noise reduction after data preprocessing.

transformation of video streams to picture representation. Such images have a definite structure of 450 × 350 pixels. The photos have been subsequently smoothened via the median processing technique. Median filtering has been conducted to recognize distorted pixels in frames and substitute these with the mean frequency. We utilized a 5 × 5 grid to eliminate noise. The mathematical formulation of the median filter is described in Eq. (1), (2), and (3):

$$Medf\ (I) = Medf\{I_m) \tag{1}$$

$$= \frac{I_m\,(n+1)}{2}; n\ is\ odd \tag{2}$$

$$\frac{1}{2}\left[I_m\left(\frac{n}{2}\right) + I_m\left(\frac{n}{2}\right) + 1\right] \tag{3}$$

where $I1, I2, I3, \ldots$, it represents the frequency of neighboring pixels. All of the image's accessible pixels has been presented in order. The succeeding identification and organization of the dependent on the applied is I $m1I\ m2I\ m3I\ mn$, whereby n has been typically anomalous. The outcomes of noise reduction and data preparation are depicted in Fig. 2.

B. SILHOUETTE EXTRACTION

After preprocessing, the next step is to extract the silhouette and background subtraction. For this, we initially applied a Gaussian mixture model and supper pixel model over RGB images and subtracted the background. Then, a super pixel combining method comparable to the one suggested by Yang et al. [47] was utilized to find the required human-object combination. This method involves merging neighboring and comparable superpixels to create larger super pixels based on similarities unless the necessary amount of super pixels has been reached. Each super pixel has four extracted features: average, autocorrelation, scale-invariant frequency extractor, and speeded-up robust features. These three principles have been concatenated to create a super pixel classification model. After that, we have converted it into binary image format. We have dealt in binary due to less computational cost and processing power. Binary images are based on (0,1), while RGB images have three matrices and a 0-255 range. Fig. 3 shows the results of silhouette extraction.

1) HUMAN DETECTION

In this section, we have discussed the procedure for human detection. We have taken background subtracted images as

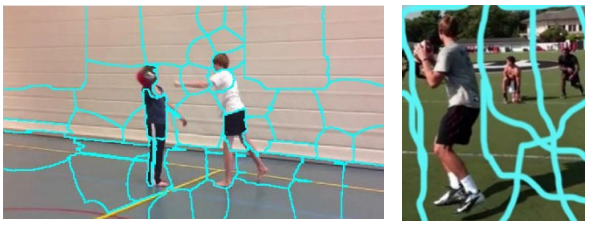


FIGURE 3. Supper pixel result over RGB images of sports video in the wild dataset.

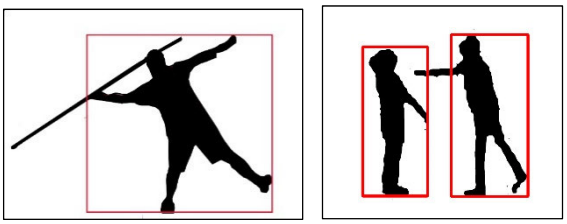


FIGURE 4. The results of human detection over extracted foreground.

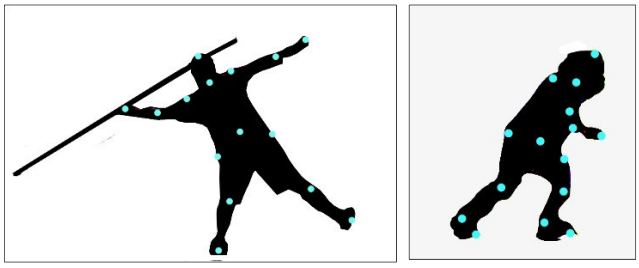


FIGURE 5. The results of human body key point's detection.

input for this step. We applied the human body shape model over-extracted human silhouette and detect the human head. Eq. (4) shows the representation of human head tracking which helped us for the detection in extracted foreground images.

$$Ih_{hp}^{f} < -Ih_{he}^{f-1} \Delta Ih_{he}^{f-1} \tag{4}$$

where Ih_{hp}^{f} head location of specified image or frame. Fig. 4 shows the subtracted background binary image results and human detection with bounding box.

2) HUMAN KEY POINTS DETECTION

We tracked the human head and used bounding box to the entire human body. Eq. (5) shows the mathematical association of human body torso points.

$$Ih_{tp}^{f} < -Ih_{t}^{f-1} \Delta Ih_{t}^{f-1} \tag{5}$$

where Ih_{tp}^{f} is the torso location of specified image or frame. Fig. 5 shows the results of human body point's recognition.

3) OBJECT DETECTION

The rapid shift method has been used to separate the provided picture into super pixels; Every two neighboring super pixels has been combined into a single large super pixel, if the resemblance between their feature trajectories exceeds a threshold. The procedure has been repeated until

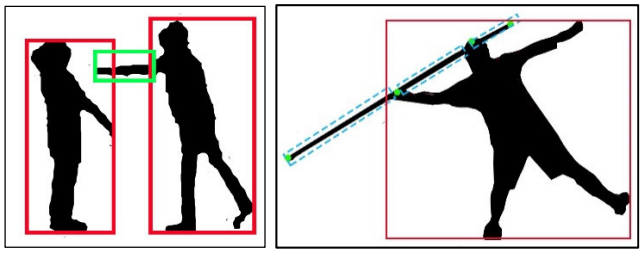


FIGURE 6. The example results of object recognition.

the backdrop, the person, and the object three super pixels remain. The super pixel with the biggest area has been deleted and assigned the backdrop to produce the desired human silhouette. The formulation has been used to determine the median (xm, ym) of any existent locations, j and k. Eq. (6).

$$(x_m, y_m) = (\frac{x_j + x_k}{2}, \frac{y_i + y_k}{2}) \tag{6}$$

where x and y are the pixel position. Fig. 6 illustrations the results of object recognition.

Algorithm 1, describes the human body parts detection technique in detail.

Algorithm 1 Human Body Key Points Detection

Input: RS: Extracted Outer shape of human body
Output: 15 body parts detected
RS=human body shape, S = human shape, H=height, W = width, L=left, R=right, HR= head, N=neck
Repeat
For*i* = 1 *to N* ***do***
Search (*RS*)
$HR = head\_point\_Area$;
$I\_H = UpperPoint\ (HR)$
$EH = End_{Head_{point}}\ (HR)$
$I\_Feet = Bottom\ (S)$
$I\_Mid = mid\ (H, W)/2$
$I\_Foot = Bottom(S)\&\ search(L, R)$
$I\_K = mid(Img\_Mid, ImgFoot)$
$I\_H = HR\&search(L, R)$
$I\_S = search(HR, N)\&search(R, L)$
$I\_E = mid(Img\_H, ImgS)$
End
Until main sections of input are examined.

C. FEATURES EXTRACTIONS

During this section, we established the mean and frame occupancy to separate the trajectory components from the supplied photos. The internal consistency approach detected elements, and the trajectory solvent system separated the critical components from those objects. The mean and frame occupancy estimates used in this investigation have been derived from identified ingredients. The assembled members made up the subset of functions. The trajectories involved in the methodology for trajectory recognition are described

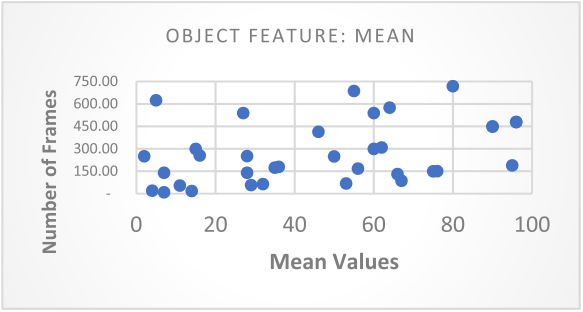


FIGURE 7. Object trajectory mean values graphical representation.

below. Algorithm 2 shows the overview of trajectories extraction.

1) OBJECT FEATURE I: MEAN

Within a series, the object's velocity varies from image sequences, and therefore the distance between the entity's places in every frame has been computed. The changed performance index for each ingredient over repeating bunches has been then estimated, and a trajectory has been determined. The histogram below illustrates the median distance between consecutive pixels,

$$P\left(A_o, A_{o+1}\right) = \frac{\sum_{o=1}^{N} C}{N}, \tag{7}$$

where C denotes the distance travelled by the particle between the oth and $(o+1th)$ the edge. Tenure N denotes the total number of buildings. Fig. 7 shows the results of object trajectories mean.

2) OBJECT FEATURE II: FRAME OCCUPANCY

This module represents the quantity of width an object consumes on each monitor. When events transpire between photos, the territory symbolized by a person on each screen changes. Identifying the length of something may define its shape, allowing for its classification. Consequently, it is vital to identify the damage induced by each individual standard:

$$F^o = \sum_{b=1}^{K} \sum_{v=1}^{L} J_{bv} \tag{8}$$

$$J_{bv} = \begin{cases} 1 & ; if\ object \\ 0 & ; otherwise \end{cases} \tag{9}$$

Fig. 8 shows the results of object trajectories frame occupancy over SVW dataset.

After this we have illustrated the procedure and method of various feature extraction over three complex datasets: R-transform, Direction cosine and angular movement. Algorithm 3 shows the overview of feature extraction.

3) SILHOUETTE FEATURE I: R-TRANSFORM

R-Transform is based on full body features of an extracted human silhouette. We only took the area of interest and applied R-transform to the human body. We got the various data related to R-transform. Finally, we found the mean value of every human body and map it in the R-Transform vector.

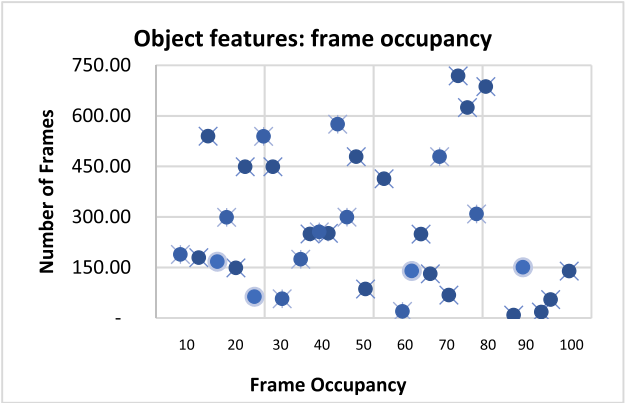


FIGURE 8. Object features frame occupancy values graphical representation.

Algorithm 2 Object Features Extraction (Ofe)

Input: Input_data
Output: Extracted Ofe vectors $(t_1, t_2, t_3 \ldots\ldots t_n)$
Extarcted_Ofe ← []
Ofe_Data ← Get_*Ofe*_Datal()
*Ofe*_Data_size← Get_*Ofe*_Data_size()
Procedure HAA(Video, Images)
OfeVector ← []
*Ofe*_Data← Object(Mean, FrameOccupancy)
Sampled_*Ofe*_Data(*Ofe*Data)
While exit void state do
[*Me*, *Fo*] ←Extractl*Ofe*(*Ofe* data)
ExtractedOfeVector ← [*Me*, *Fo*]
Return *Ofe*_Vector

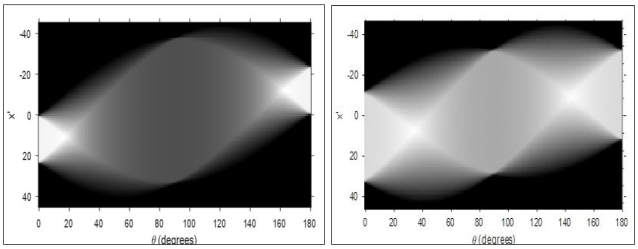


FIGURE 9. The example results of R-Transform features over human extracted body.

The R-transform $R(\rho, \theta)$ of given silhouette $S(x, y)$ has been be achieved as follows:

$$R\left(p, \theta\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y)\vartheta(p - xcos\theta - ysin\theta)dxdy \tag{10}$$

Fig. 9 demonstrates the results of R-transform features.

4) SILHOUETTE FEATURE II: DIRECTION COSINE

In this feature value, we considered the human body's undertaking direction and applied the cosine formula over the extracted movement. Direction values and angular values of cosine have been regarded as features. We used this procedure

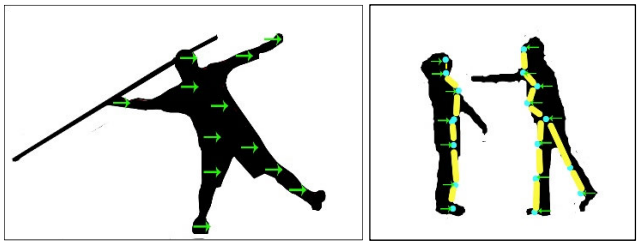


FIGURE 10. The example results of direction cosine (a) indicates the direction and (b) shows the relationship of direction cosine.

over all the humans in a given frame and the entire dataset—finally mapped the direction cosine vector. Eq. (11) shows the formulation of direction cosine features.

$$d = \sum_{o}^{p} i\,(x, y) \rightarrow a \tag{11}$$

where d is the direction flow of extracted silhouette x, y is the directory standards of the given frame, i is RGB (x,y,z) pixel values, and $\rightarrow$ d illustrates the gesture. To find the cosine vales we used the cosine formula as;

$$sin^2 d = cos^2 d - 1 \tag{12}$$

where d indicates the directional values and sin/cos shows the angular values. Fig. 10 shows the results of direction cosine features.

5) SILHOUETTE FEATURE III: ANGULAR MOVEMENT

Angular movement features are usually based on motion information of the human body and its approach to following motion information of the same body. Initially, we applied the motion information extraction method via change detection. Now we considered them as $M1, M2 \ldots .Mn$ after this we used angular movement over $m1$ *and* $m2$ and found the angle between them; next $m2$ *and* $m3$ have been considered the next angle. The same procedure has been applied over all the extracted points. The Eq. (13) shows the measured relation of angular movement features.

$$\begin{aligned} &M1 = tan\,(X + Y) \rightarrow d1, M2 = tan\,(X + Y) \rightarrow d2 \\ &M3 = tan\,(X + Y) \rightarrow d3, M4 = tan(X + Y) \rightarrow d4 \\ &Mn = tan(X + Y) \rightarrow dn \\ &While\ tan(X + Y) = tanX + tanY/1 - tanX\ tanY \end{aligned} \tag{13}$$

where $m1 - mn$ shows the angular movement, cos denotes the values and $d1 - dn$ shows the distance. Fig. 11 shows the results of angular movement features.

D. STOCHASTIC GRADIENT DECENT

Gradient Descent optimization approaches created a set of procedures that evaluate the efficient learning pattern in order to discover the optimal feature arrangement with the cheapest possible portion. However, the stochastic gradient approach performed slowly when assessing all classification techniques from each phase and when the quantity of training variables was substantial. The Sequential Stochastic

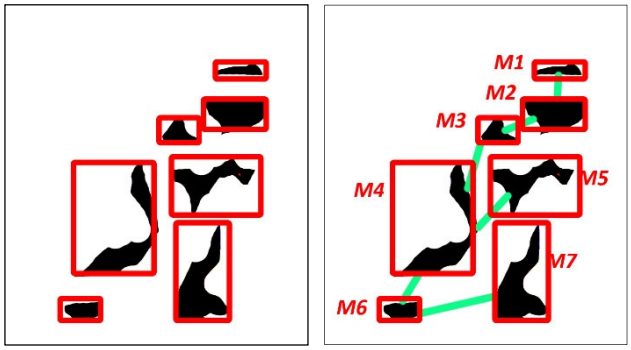


FIGURE 11. The example results of angular movement features (a) shows the moveable body parts while (b) indicates the label and angle between two points.

Algorithm 3 Feature Calculation

Input: Input_data
Output: Extracted Feature vectors ($f_1, f_2, f_3 \ldots\ldots f_n$)
Extarcted_features ← []
F_Data ← Get_F_Datal()
F_Data_size ← Get_F_Data_size()
Procedure (Video, Images)
FeaturesVector ← []
Denoise_F_Data← Pre_processing(Win,Median)
Sampled_F_Data(DenoiseData)
While exit void state do
$[RT, DC, AMF]$ ←ExtractlFeatures(sample data)
ExtractedFeaturesVector ←$[RT, DC, AMF]$
Return Context_intelligent_features_Vector

Gradient method (SGD) with a representative sample has been proposed as an optimization procedure that does not employ all preparation stages to solve the issue. Furthermore, not a single piece of data was processed. Minibatch SGD randomly selects a subset of data to reduce costs and obtain a long wavelength than standard SGD. Minibatch requires continual assessment, effective learning patterns, and beginning factors to design a performance index with a potential for a modest loss. However, the background learning rate 0.01 has been changed by optimizing the regularization backpropagation using the Leave One Subject Out (LOSO) merging. The SGD of all optimization models for x (k) and y (k) labels is approximated:

$$\theta = \theta - n.\nabla_\theta k(\theta; x^{(k)}; y^{(k)}) \tag{14}$$

where nbs being the minibatch dimension, it achieves a decreased variance of starting characteristics and smaller back propagation. Fig. 12 shows the results of optimized features vector via the stochastic gradient descent algorithm.

E. RESTRICTED BOLTZMANN MACHINES

It is a neural network with each cell interconnected to each other synapse. This mechanism has multiple pairs: the apparent surface or number of neurons and the concealed

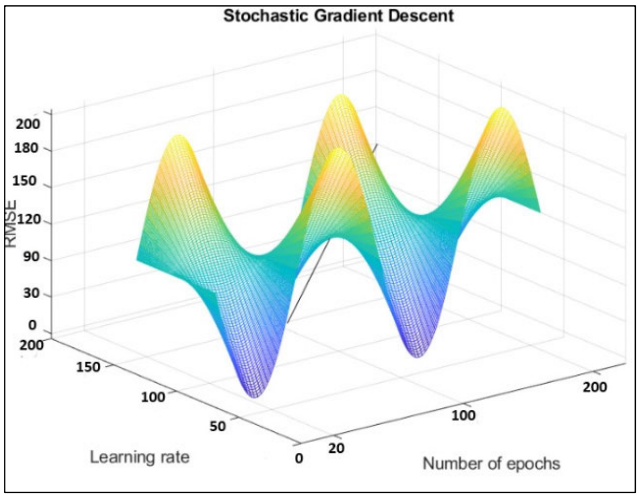


FIGURE 12. The results of optimized features vector via stochastic gradient descent algorithm.

layer. The visible layer has been represented by v and the concealed layer by h. The Boltzmann machine lacks an objective function. Boltzmann machines are unpredictable and creative neural networks accomplished by developing embedding and representing and (given sufficient time) cracking challenging innovation and entrepreneurship issues. The Boltzmann distribution (also called Gibbs probability) is an essential component of mathematical analysis. It explains the effect of factors such as concentration and temperatures on subatomic particles in thermodynamics. Consequently, it is sometimes referred to as the Energy-Based Approach. Eq. (15)-(19) shows RBM's relation and mathematical formulation.

$$P\left(h_{1j}=1|\upsilon_1\right)=g\left(b_j+\sum\nolimits_i \upsilon_{1i}W_{ij}\right) \tag{15}$$

$$P\left(\upsilon_{2i}=1|h_1\right)=g\left(c_i+\sum\nolimits_j W_{ij}h_{1j}\right) \tag{16}$$

$$P\left(h_{2j}=1|\upsilon_2\right)=g\left(b_j+\sum\nolimits_i \upsilon_{2i}W_{ij}\right) \tag{17}$$

$$\begin{aligned} W &= W+\eta\left(h_1\upsilon_1-h_2\upsilon_2\right); \\ c &= c+\eta\left(\upsilon_1-\upsilon_2\right); \\ b &= b+\eta\left(h_1-h_2\right) \end{aligned} \tag{18}$$

$$\begin{aligned} H_j &= g\left(b_j+\sum\nolimits_i \upsilon_{1i}W_{ij}\right), \\ V_i &= g\left(c_i+\sum\nolimits_j W_{ij}h_{1j}\right) \end{aligned} \tag{19}$$

where spectrum vector $v_1 = \left(v_{11},\ldots,v_{1n,}\right)$ (with size n_u), the size of the hidden layer n_h, the learning rate η, and the maximum epochM_e $\mathrm{H} = \left(H_1,\ldots,H_{n_h}\right)$ (a hidden vector); $\mathrm{V} = \left(V_1,\ldots,V_{n_v}\right)$ (reconstruction of spectrum υ). Fig. 13 shows the conceptual model of RBM.

IV. EXPERIMENTAL ANALYSIS AND RESULTS

The leave-one-subject-out (LOSO) cross-validation approach was used to evaluate the performance of the proposed methodology on three publicly accessible resources, including the MPII pose track, the UCF aerial action collection,

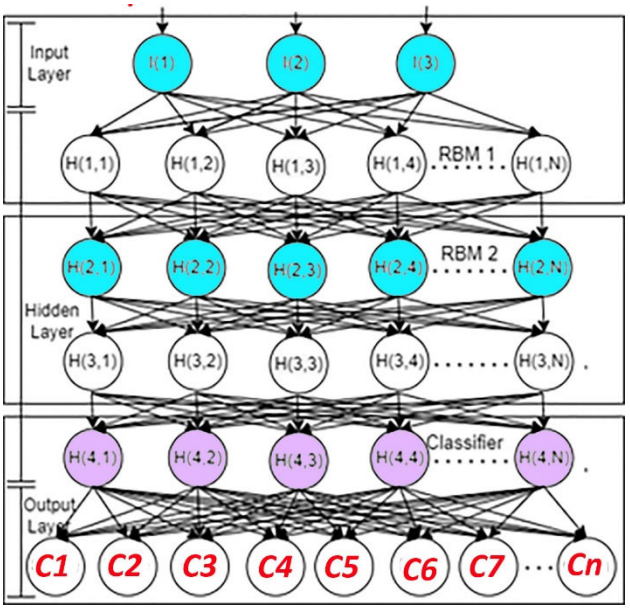


FIGURE 13. The conceptual model of RBM.

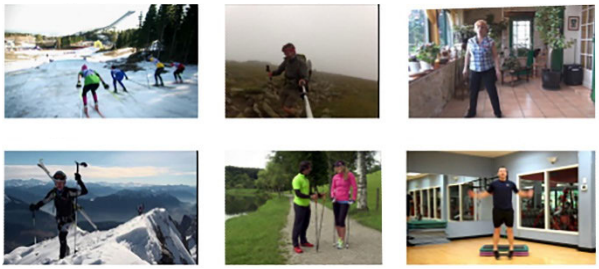


FIGURE 14. Example images from various classes of the MPII human pose dataset.

and the Sports Videos in the Wild (SVW) dataset. The LOSO method is an enhanced cross-validation approach that utilizes data from a single participant for each fold.

The MPII Human Pose dataset is a standard for evaluating multimodal human pose computation. The collection consists of around 25K photos of over 40K individuals with documented sequences. The photos were captured methodically using a predefined categorization of commonplace human actions. The collection includes 410 human interactions, and each visual is tagged with a corresponding activity. Each picture was pulled from a YouTube video and accompanied by unlabeled data frames before and after it. In addition, we got richer descriptions for the test set, such as body component geometric distortion and 3D torso and head orientation. Figure 14 illustrates a selection of photographs from the MPII human posture collection.

This data collection contains image sequences captured with an R/C-controlled blimp carrying a Video camera placed on a reference plane. The accumulation has various airborne and ground-level movements at varying heights. Several occurrences of each movement were captured at various altitudes ranging from 400 to 450 meters and accomplished

FIGURE 15. Example images from various classes of the UCF aerial action dataset.

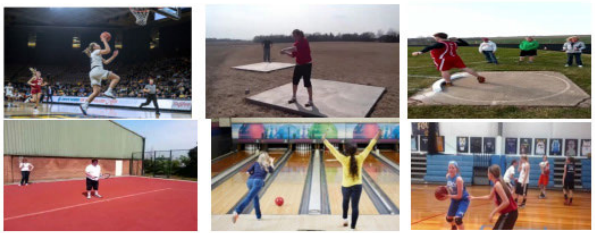


FIGURE 16. Example images from various classes of the SVW dataset.

by various characters. Fig. 15 illustrates a sampling of the photos from the UCF aerial action dataset.

The Sports videos in the wild dataset contain videos of sportspersons practicing dissimilar sports. We have attained all video sequences from YouTube and explained their class label with the assistance of Amazon Mechanical Turk. Fig. 16 illustrates a sampling of the photos from the SVW dataset.

A. HARDWARE PLATFORM

In the experimental setup of the designed approach, we used MATLAB (R2021a) for all simulations and analysis over Intel (R) Core (TM) i7-8665U @ 1.90GHz CPU with 64-bit Windows 11 in the testing device. The device had 16 GB of RAM. The new finding on the MPII pose Dataset, UCF aerial dataset and SVW datasets along with experimental conclusions, is examined in the outcomes segment.

B. EXPERIMENTAL SETUP AND EVALUATION

We calculated the Euclidean distance between the actual data and considered points using the Euclidean spectrum to evaluate the effectiveness of the human body key points detection structure. The minimum level distance of the actual data at which the failure ratio is placed to 13 was used to determine the average accuracy.

$$Ed_d = \sqrt{\sum_{n=1}^{n} \left(\frac{X_g}{S_g} - \frac{X_d}{S_d} \right)^2} \tag{20}$$

Y_d is an identified position of the proposed approach, X_g is the characterized data set, and Ed_d is the Euclidean distance.

TABLE 1. Human body key point's detection and distance details over the MPII, UCF aerial and SVW dataset.

Key points	MPII Dataset		UCF_A Dataset		SVW Dataset	
	Distance	Accuracy (%)	Distance	Accuracy(%)	Distance	Accuracy(%)
Head	11.1	91	12.3	80	10.9	91
Neck	11.3	90	12.5	79	11.1	92
R. Sh	10.2	89	12.1	79	10.7	89
L. Sh	9.5	90	11.2	80	11.2	90
R. El	9.5	91	12.4	85	10.3	90
L.EL	9.7	88	12.9	88	11.7	89
R.Hd	8.9	87	10.2	87	11.1	86
L.Hd	9.9	88	11.2	86	10.3	88
Mid	10.3	87	11.4	87	11.1	90
R.Kn	10.2	90	12.1	86	10.4	91
L. Kn	11.1	88	11.7	79	10.2	87
R. Fe	10.4	87	11.4	77	11.1	84
L.Fe	10.5	84	11.7	73	10.3	81
Mean	11.54	88.46	18.00	82.00	11.70	88.30

R.sh = right shoulder, L.Sh= left shoulder, R.El is right elbow, L.El, left elbow, R.Hd = right hand, L.Hd = left hand, R.Kn= Right knee, L.Kn= left knee, R.Fe= right foot, L.Fe= left foot.

We utilized Eq. (21) to estimate body parts' validity.

$$A_{dp} = 100/n[\sum_{n=1}^{n} \begin{cases} 1, if\ Ed_d \leq 13 \\ 0, Ed_d > 13 \end{cases}] \tag{21}$$

where A_{dp} symbolizes the estimated precision of N parts of the body, if the approximate displacement of an identified human key point was greater than 14, it was disregarded. Alternatively, the identified key point of the human body was considered in the assessment method. We achieved 88.46% of human body key point's recognition accuracy over MPII dataset, 82.00% of human body key point's recognition accuracy over the UCF aerial action dataset and 88.30% of human body key points recognition accuracy over the SVW dataset. Table 1 shows the detailed results and error rate over three datasets.

The next step is to find the mean accuracy of our proposed structure, we applied a robust classification algorithm namely Restricted Boltzmann Machine. Table 2 presents the confusion matrix for the MPII pose track dataset with an accuracy rate of 92.71%. Table 3 illustrates the confusion matrix for the UCF aerial action dataset, which has an average accuracy of 90.60%. Table 4 shows the confusion matrix for the SVW dataset, which has an average accuracy of 92.42%.

C. COMPARISON WITH OTHER STATE-OF-THE-ART SYSTEMS

After this we compared our suggested approach with other methods Yang et al [47]'s suggestion to create a bypass connection to a deep network was motivated by the ResNet remainder block's architectural design. When all remnant blocks have been changed with a sparsely linked residual

TABLE 2. Confusion matrix of MPII dataset.

Interactions	ML	HR	SB	BS	RS	RD	GL
ML	93	0	0	2	0	1	0
HR	0	91	0	2	0	0	0
SB	0	2	92	1	1	0	0
BS	0	0	0	94	1	0	0
RS	1	0	1	0	93	1	0
RD	2	0	1	2	2	94	0
GL	0	0	0	2	0	1	92
Mean accuracy rate = 92.71%							

* ML= moving lawn, HR= horseback riding, SB=skateboarding, BS=Bicycling, RS= rope skipping, RD= ride surfboard, GL = golf

TABLE 3. Confusion matrix of UCF aerial action dataset.

Interactions	PO	OD	CR	OT	CT
PO	92	0	0	0	2
OD	0	90	2	0	0
CR	0	2	90	0	0
OT	1	0	1	91	1
CT	2	0	0	0	90
Mean accuracy rate = 90.60%					

* PO = picking up an object, OD = open a car door, CR = closing a car door, OT = Opening a car trunk, CT= closing a car trunk

TABLE 4. Confusion matrix of SVW dataset.

Interactions	AR	BB	BT	BW	FB	BO	SK
AR	95	0	0	2	0	1	0
BB	1	92	0	0	0	0	0
BT	0	1	91	1	1	0	0
BW	0	0	0	90	1	1	0
FB	1	0	1	0	94	1	0
BO	2	0	1	2	2	92	1
SK	0	2	2	2	0	1	93
Mean accuracy rate = 92.42%							

* AR = archery, BB = baseball, BT = basketball, BW = bowling, FB = football, BO=Bowling, SK = skating

component, the required number of constraints has been reduced dramatically, and the computational time has been lowered, increasing the cable network divergence. Using a methodology, Zhang et al. [48] explained how to estimate human location. By using a spatially separated recurrent neural networks search strategy, they lowered calculation time with practically any performance loss while using the MobileNetV2 internet backbone architecture for attitude determination. Wang et al. [49] developed an additional essentially lightweight bottleneck in UULPN with a large variety of mappings and a wide range of them. Researchers

TABLE 5. Human and object interaction classification comparison of recognition rate of the proposed method with other state-of-the-art methods over MPII, UCF aerial action and SVW datasets.

Methods	MPII (%)	UFC_A (%)	SVW (%)
Zhang et al. [48]	88.1%	--	--
Yang et al. [47]	88.8%	--	--
Wang et al. [49]	85.7%	78.1%	--
S. Sun et al. [50]	--	--	74.2%
Reza. F et al. [51]	--	79.3%	82.3%
Zhu, Y [52]	--	--	83.1%
Agwad, El [53]	--	80.0%	--
Li et al[54]		78.22%	
H. W. Chen et al[55]		79.0%	
Rodriguez et al [56]	71.80%		
Proposed	**92.71**	**90.60**	**92.42**

provided both a multi and single-branch construction for the bottleneck's region. To increase prediction performance, a cross-topology was implemented during the training process. Sun et al. in [50] suggested a method for extracting features in which they retrieve focused motion information in addition to a CNN-based model for classifying and identifying human events. Reza. F. et al. established a strategy in [51] for dealing with episode identification and classification using CNN and Connection in Network architecture (NNA), which forms the foundation of contemporary CNN. The minimalist Convolution layer, averaged, max, and product characteristics are employed to recognize human interactions. Researchers created a human locomotion approximation algorithm in [52], where they used a video-based device to enhance dense characteristics. Kinematic Regularization and Matrix Rank Optimization (KRMARO) is a revolutionary strategy proposed in [53] study to obtain high true-detection percentages and drastically reduced falsified rate increases. KRMARO describes the problem of identifying moving objects by combining a novel kinematic regularization with the exploratory factor chase. They take optical flow into account for the multi - faceted material. To expedite model inference during the build process, a separate design was used. Li and Chan [54] use conventional neural networks for determining human body posture. Chen and McGurr [55] used morphology separation and systematic the thresholding to determine the surface color. Rodriguez et al [56]. presented an innovative technique for predicting future body motion. Reasonable justifications and targeted failure procedures were used to encourage reproductive systems to predict specified future human motion. The assessment of human-object interaction and classification with state-of-the-art techniques is presented in Table 5.

Next step is to find precision, recall and F-1 score using formula, In Tables 6, 7 and 8 we estimate the precision, recall

TABLE 6. Measurements of evaluation metrics for the proposed system over the MPII dataset.

MPII	AdaBoost			ANN			RBM		
Classes	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
ML	0.945	0.935	0.940	0.947	0.967	0.957	0.969	0.969	0.969
HR	0.906	0.926	0.916	0.956	0.967	0.961	0.978	0.978	0.978
SB	0.912	0.943	0.927	0.956	0.946	0.951	0.979	0.958	0.968
BS	0.913	0.913	0.913	0.914	0.934	0.924	0.913	0.989	0.949
RS	0.933	0.923	0.928	0.957	0.967	0.962	0.959	0.969	0.964
RD	0.935	0.888	0.911	0.947	0.918	0.933	0.969	0.931	0.949
GL	0.914	0.934	0.924	0.978	0.957	0.967	0.979	0.968	0.984

(Note: ML= moving lawn, HR= horseback riding, SB=skateboarding, BS=Bicycling, RS= rope skipping, RD= ride surfboard, GL = golf).

TABLE 7. Measurements of evaluation metrics of the proposed system over the UCF aerial dataset.

UCF_A	AdaBoost			ANN			RBM		
Activities	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
PO	0.924	0.977	0.950	0.968	0.978	0.973	0.968	0.979	0.974
OD	0.903	0.923	0.913	0.955	0.955	0.955	0.978	0.978	0.978
CR	0.922	0.933	0.927	0.944	0.933	0.939	0.968	0.978	0.973
OT	0.941	0.899	0.920	0.955	0.955	0.955	0.968	0.968	0.984
CT	0.908	0.919	0.913	0.934	0.934	0.934	0.968	0.978	0.973

(Note: PO = picking up an object, OD = open a car door, CR = closing a car door, OT = Opening a car trunk, CT= closing a car trunk).

TABLE 8. Measurements of evaluation metrics for the proposed system over the SVW dataset.

SVW	AdaBoost			ANN			RBM		
Classes	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
AR	0.937	0.989	0.962	0.968	0.989	0.978	0.960	0.990	0.974
BB	0.947	0.957	0.952	0.978	0.968	0.973	0.968	0.989	0.979
BT	0.956	0.946	0.951	0.968	0.957	0.963	0.958	0.968	0.963
BW	0.935	0.916	0.926	0.957	0.978	0.967	0.947	0.978	0.963
FB	0.944	0.924	0.934	0.957	0.967	0.962	0.959	0.969	0.964
BO	0.926	0.916	0.921	0.947	0.938	0.942	0.958	0.920	0.939
SK	0.955	0.955	0.955	0.978	0.957	0.967	0.989	0.930	0.959

(Note: PO = picking up an object, OD = open a car door, CR = closing a car door, OT = Opening a car trunk, CT= closing a car trunk).

and F-1 score or F-measure using Adabosst, ANN and RBM classification algorithms.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{22}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{23}$$

$$F1 - Score = \frac{2(Precision \times Recall)}{(Precision + Recall)} \tag{24}$$

We compared their precision, their recall, and the F1 scores of all classes used in the three benchmark datasets, MPII pose, UCF aerial, and SVW dataset.

V. DISCUSSION

This paper is based on human-object interaction in the human natural living environment. We can apply this approach and framework in various public environments such as entertainment, sports, technology, education, intelligent surveillance system, and innovative emergency services. We applied this technique in three publicly available datasets and achieved an intelligent amount of accuracy with a low error rate, a 7.21% error rate for the MPII pose dataset, a rate of 9.40% for UCF, and a 7.38% of error rate for SVW dataset. Due to the complex nature of these benchmark datasets, this study has one drawback: occlusion and shadow effects. This problem impacted human tracking and verification and

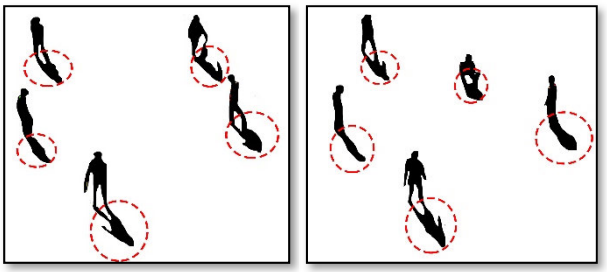


FIGURE 17. Limitation example images of proposed method.

the feature engineering process. This is the main factor that caused the mean recognition to drop. Fig. 17 shows the limitation example of the proposed method.

VI. CONCLUSION AND FUTURE WORK

In this paper, we designed a framework for human and object interaction in real-life examples such as sports and other activities. As a result, we achieved a butter accuracy of human body parts detection for the MPII dataset is 88.46%. For the UCF_aerial Dataset, it is 82.00% and for sports video in the wild Dataset, it is 88.30%. The classification accuracy for the MPII dataset is 92.71%, 90.60% for the UCF_aerial dataset, and 92.42% for the sports video in the wild Dataset. In future work, we will integrate more composite tasks from various contexts, including medical centers, workplaces, and smart homes. We will also fuse more feature engineering techniques from different domains to recognize complex motion patterns in multiple contexts.

REFERENCES

[1] M. Mahmood, 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 Appl.*, vol. 79, nos. 11–12, pp. 6919–6950, Mar. 2020.

[2] N. Naz, M. K. Ehsan, M. R. Amirzada, M. Y. Ali, and M. A. Qureshi, "Intelligence of autonomous vehicles: A concise revisit," *J. Sensors*, vol. 2022, pp. 1–11, Apr. 2022.

[3] X. Zhou and L. Zhang, "SA-FPN: An effective feature pyramid network for crowded human detection," *Int. J. Speech Technol.*, vol. 52, no. 11, pp. 12556–12568, Sep. 2022.

[4] A. Jalal, N. Sarif, J. T. Kim, and T.-S. Kim, "Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home," *Indoor Built Environ.*, vol. 22, no. 1, pp. 271–279, Feb. 2013.

[5] A. Jalal, Y. Kim, and D. Kim, "Ridge body parts features for human pose estimation and recognition from RGB-D video data," in *Proc. 5th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2014, pp. 1–6.

[6] M. A. K. Quaid and A. Jalal, "Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm," *Multimedia Tools Appl.*, vol. 79, nos. 9–10, pp. 6061–6083, Mar. 2020.

[7] A. Jalal and Y. Kim, "Dense depth maps-based human pose tracking and recognition in dynamic scenes using ridge data," in *Proc. 11th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2014, pp. 119–124.

[8] B. Li, M. Zhang, Y. Rong, and Z. Han, "Transceiver optimization for wireless powered time-division duplex MU-MIMO systems: Non-robust and robust designs," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4594–4607, Jun. 2022.

[9] A. Nadeem, A. Jalal, and K. Kim, "Human actions tracking and recognition based on body parts detection via artificial neural network," in *Proc. 3rd Int. Conf. Advancements Comput. Sci. (ICACS)*, Feb. 2020, pp. 1–6.

[10] F. Wang, H. Wang, X. Zhou, and R. Fu, "A driving fatigue feature detection method based on multifractal theory," *IEEE Sensors J.*, vol. 22, no. 19, pp. 19046–19059, Oct. 2022.

[11] M. Batool, A. Jalal, and K. Kim, "Telemonitoring of daily activity using accelerometer and gyroscope in smart home environments," *J. Electr. Eng. Technol.*, vol. 15, no. 6, pp. 2801–2809, Nov. 2020.

[12] L. Yan, Y. Shi, M. Wei, and Y. Wu, "Multi-feature fusing local directional ternary pattern for facial expressions signal recognition based on video communication system," *Alexandria Eng. J.*, vol. 63, pp. 307–320, Jan. 2023.

[13] M. Pervaiz, A. Jalal, and K. Kim, "Hybrid algorithm for multi people counting and tracking for smart surveillance," in *Proc. Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2021, pp. 530–535.

[14] T. Li, Y. Fan, Y. Li, S. Tarkoma, and P. Hui, "Understanding the long-term evolution of mobile app usage," *IEEE Trans. Mobile Comput.*, vol. 22, no. 2, pp. 1213–1230, Feb. 2023.

[15] A. Jalal, M. Batool, and K. Kim, "Sustainable wearable system: Human behavior modeling for life-logging activities using K-Ary tree hashing classifier," *Sustainability*, vol. 12, no. 24, p. 10324, Dec. 2020.

[16] Z. Liu, P. Qian, J. Yang, L. Liu, X. Xu, Q. He, and X. Zhang, "Rethinking smart contract fuzzing: Fuzzing with invocation ordering and important branch revisiting," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1237–1251, 2023.

[17] A. Arif and A. Jalal, "Automated body parts estimation and detection using salient maps and Gaussian matrix model," in *Proc. Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2021, pp. 667–672.

[18] J. Luo, Y. Wang, and G. Li, "The innovation effect of administrative hierarchy on intercity connection: The machine learning of twin cities," *J. Innov. Knowl.*, vol. 8, no. 1, Jan. 2023, Art. no. 100293.

[19] A. Ahmed, A. Jalal, and K. Kim, "Multi-objects detection and segmentation for scene understanding based on texton forest and kernel sliding perceptron," *J. Electr. Eng. Technol.*, vol. 16, no. 2, pp. 1143–1150, Mar. 2021.

[20] A. Nadeem, A. Jalal, and K. Kim, "Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy Markov model," *Multimedia Tools Appl.*, vol. 80, no. 14, pp. 21465–21498, Jun. 2021.

[21] A. Jalal, A. Ahmed, A. A. Rafique, and K. Kim, "Scene semantic recognition based on modified fuzzy C-mean and maximum entropy using object-to-object relations," *IEEE Access*, vol. 9, pp. 27758–27772, 2021.

[22] S. Xiong, B. Li, and S. Zhu, "DCGNN: A single-stage 3D object detection network based on density clustering and graph neural network," *Complex Intell. Syst.*, vol. 9, no. 3, pp. 3399–3408, Jun. 2023.

[23] C. Fu, H. Yuan, H. Xu, H. Zhang, and L. Shen, "TMSO-Net: Texture adaptive multi-scale observation for light field image depth estimation," *J. Vis. Commun. Image Represent.*, vol. 90, Feb. 2023, Art. no. 103731.

[24] A. A. Rafique, A. Jalal, and K. Kim, "Automated sustainable multi-object segmentation and recognition via modified sampling consensus and kernel sliding perceptron," *Symmetry*, vol. 12, no. 11, p. 1928, Nov. 2020.

[25] D. Li, S. S. Ge, and T. H. Lee, "Fixed-Time-Synchronized consensus control of multiagent systems," *IEEE Trans. Control Netw. Syst.*, vol. 8, no. 1, pp. 89–98, Mar. 2021.

[26] Z. Lv, Z. Yu, S. Xie, and A. Alamri, "Deep learning-based smart predictive evaluation for interactive multimedia-enabled smart healthcare," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 1s, pp. 1–20, Feb. 2022.

[27] M. Mahmood, A. Jalal, and M. A. Sidduqi, "Robust spatio-temporal features for human interaction recognition via artificial neural network," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2018, pp. 218–223.

[28] A. Jalal, M. A. K. Quaid, and M. A. Sidduqi, "A triaxial acceleration-based human motion detection for ambient smart home system," in *Proc. 16th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2019, pp. 353–358.

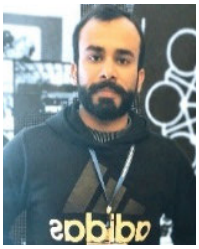
[29] S. Haresh, X. Sun, H. Jiang, A. X. Chang, and M. Savva, "Articulated 3D human-object interactions from RGB videos: An empirical analysis of approaches and challenges," 2022, *arXiv:2209.05612*.

[30] M. Pervaiz and A. Jalal, "Artificial neural network for human object interaction system over aerial images," in *Proc. 4th Int. Conf. Advancements Comput. Sci. (ICACS)*, Feb. 2023, pp. 1–6.

[31] A. Jalal and M. Mahmood, "Students' behavior mining in E-learning environment using cognitive processes with information technologies," *Educ. Inf. Technol.*, vol. 24, pp. 2797–2821, Mar. 2019.

[32] U. Azmat and A. Jalal, "Smartphone inertial sensors for human locomotion activity recognition based on template matching and codebook generation," in *Proc. Int. Conf. Commun. Technol. (ComTech)*, Sep. 2021, pp. 109–114.

[33] Z. Su, L. Xu, D. Zhong, Z. Li, F. Deng, S. Quan, and L. Fang, "RobustFusion: Robust volumetric performance reconstruction under human-object interactions from monocular RGBD stream," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6196–6213, May 2023.
[34] M. Muneeb, H. Rustam, and A. Jalal, "Automate appliances via gestures recognition for elderly living assistance," in *Proc. 4th Int. Conf. Advancements Comput. Sci. (ICACS)*, Feb. 2023, pp. 1–6.
[35] Q. Wang, J. Hu, Y. Wu, and Y. Zhao, "Output synchronization of wide-area heterogeneous multi-agent systems over intermittent clustered networks," *Inf. Sci.*, vol. 619, pp. 263–275, Jan. 2023.
[36] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8359–8367.
[37] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 401–417.
[38] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[39] M. Khan, M. Schneider, M. Farid, and M. Grzegorzek, "Detection of infantile movement disorders in video data using deformable part-based model," *Sensors*, vol. 18, no. 10, p. 3202, Sep. 2018.
[40] M. H. Khan, J. Helsper, M. S. Farid, and M. Grzegorzek, "A computer vision-based system for monitoring vojta therapy," *Int. J. Med. Informat.*, vol. 113, pp. 85–95, May 2018.
[41] A. Jalal, N. Khalid, and K. Kim, "Automatic recognition of human interaction via hybrid descriptors and maximum entropy Markov model using depth sensors," *Entropy*, vol. 22, no. 8, p. 817, Jul. 2020.
[42] L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1568–1576.
[43] W. Yan, Y. Gao, and Q. Liu, "Human-object interaction recognition using multitask neural network," in *Proc. 3rd Int. Symp. Auto. Syst. (ISAS)*, May 2019, pp. 323–328.
[44] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4115–4124.
[45] Y.-L. Li, X. Liu, H. Lu, S. Wang, J. Liu, J. Li, and C. Lu, "Detailed 2D-3D joint representation for human-object interaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10163–10172.
[46] A. Jalal, M. Batool, and K. Kim, "Stochastic recognition of physical activity and healthcare using tri-axial inertial wearable sensors," *Appl. Sci.*, vol. 10, no. 20, p. 7122, Oct. 2020.
[47] L. Yang, Y. Qin, and X. Zhang, "Lightweight densely connected residual network for human pose estimation," *J. Real-Time Image Process.*, vol. 18, no. 3, pp. 825–837, Jun. 2021.
[48] W. Zhang, J. Fang, X. Wang, and W. Liu, "EfficientPose: Efficient human pose estimation with neural architecture search," *Comput. Vis. Media*, vol. 7, no. 3, pp. 335–347, Sep. 2021.
[49] W. Wang, K. Zhang, H. Ren, D. Wei, Y. Gao, and J. Liu, "UULPN: An ultra-lightweight network for human pose estimation based on unbiased data processing," *Neurocomputing*, vol. 480, pp. 220–233, Apr. 2022.
[50] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1390–1399.
[51] R. F. Rachmadi, K. Uchimura, and G. Koutaki, "Combined convolutional neural network for event recognition," in *Proc. Korea-Japan Joint Workshop Frontiers Comput. Vis.*, 2016, pp. 85–90.
[52] Y. Zhu, K. Zhou, M. Wang, Y. Zhao, and Z. Zhao, "A comprehensive solution for detecting events in complex surveillance videos," *Multimedia Tools Appl.*, vol. 78, no. 1, pp. 817–838, Jan. 2019.
[53] A. ElTantawy and M. S. Shehata, "KRMARO: Aerial detection of small-size ground moving objects using kinematic regularization and matrix rank optimization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 6, pp. 1672–1686, Jun. 2019.
[54] S. Li and A. B. Chan, "3D human pose estimation from monocular images with deep convolutional neural network," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 332–347.
[55] H.-W. Chen and M. McGurr, "Moving human full body and body parts detection, tracking, and applications on human activity estimation, walking pattern and face recognition," *Proc. SPIE*, vol. 9844, pp. 213–246, May 2016.
[56] C. Rodriguez, B. Fernando, and H. Li, "Action anticipation by predicting future dynamic images," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops* (Lecture Notes in Computer Science), 2019, pp. 89–105.



ISRAR AKHTER received the B.S. degree in computer science from Hamdard University, and the M.S. degree in computer science from Air University, Islamabad. He is currently a Lecturer with the Computer Science Department, Bahria University, Islamabad, and the Ph.D. Scholar with Air University. His research interests include multimedia content, artificial intelligence, machine learning, image processing, data optimization, and gait analysis.

NAIF AL MUDAWI received the master's degree in computer science from Australian La Trobe University, in 2011, and the Ph.D. degree from the College of Engineering and Informatics, University of Sussex, Brighton, U.K., in 2018. He is currently an Assistant Professor with the Department of Computer Science and Information System, Najran University. He has many published research and scientific papers in many prestigious journals in various disciplines of computer science. He was a member of the Australian Computer Science Committee.

BAYAN IBRAHIMM ALABDULLAH received the Ph.D. degree in informatics from the University of Sussex, Brighton, U.K., in May 2022. She was an Assistant Professor with the Department of Information System, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University. She teaches several courses with the Information System Department, such as data governance, system security, and database system. Her research interests include machine learning, data science, and privacy and security.

MOHAMMED ALONAZI received the B.Sc. degree from King Saud University, Saudi Arabia, in 2008, the M.Sc. degree in computer science from the Florida Institute of Technology, Melbourne, FL, USA, in 2015, and the Ph.D. degree in informatics from the University of Sussex, U.K., in 2019. He is currently an Assistant Professor with the Department of Information Systems, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia. His research interests include human–computer interaction, UX/UI, digital transformation, cyber security, and machine learning.

JEONGMIN PARK received the Ph.D. degree from the College of Information and Communication Engineering, Sungkyunkwan University, in 2009. He is currently an Associate Professor with the Department of Computer Engineering, Tech University of Korea, South Korea. Before joining the Tech University of Korea, in 2014, he was a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI), and a Research Professor with Sungkyunkwan University, South Korea. His research interests include high-reliable autonomic computing mechanism and human-oriented interaction systems.