IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 19, 2021, accepted July 22, 2021, date of publication July 30, 2021, date of current version August 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3101716

Semantic Recognition of Human-Object Interactions via Gaussian-Based Elliptical Modeling and Pixel-Level Labeling

NIDA KHALID[1], YAZEED YASIN GHADI[2], MUNKHJARGAL GOCHOO[3], (Member, IEEE), AHMAD JALAL[1], AND KIBUM KIM[4], (Member, IEEE)

[1]Department of Computer Science, Air University, Islamabad 44000, Pakistan
[2]Department of Computer Science and Software Engineering, Al Ain University, Abu Dhabi, United Arab Emirates
[3]Department of Computer Science and Software Engineering, United Arab Emirates University, Al Ain, United Arab Emirates
[4]Department of Human-Computer Interaction, Hanyang University, Ansan 15588, South Korea

Corresponding author: Kibum Kim (kibum@hanyang.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) under Grant 2018R1D1A1A02085645, in part by Korea Medical Device Development Fund Grant through Korean Government (the Ministry of Science and ICT; the Ministry of Trade, Industry and Energy; the Ministry of Health and Welfare; and the Ministry of Food and Drug Safety) under Grant 202012D05-02, and in part by Hanyang University under Grant 201800000000647.

ABSTRACT Human-Object Interaction (HOI) recognition, due to its significance in many computer vision-based applications, requires in-depth and meaningful details from image sequences. Incorporating semantics in scene understanding has led to a deep understanding of human-centric actions. Therefore, in this research work, we propose a semantic HOI recognition system based on multi-vision sensors. In the proposed system, the de-noised RGB and depth images, via Bilateral Filtering (BLF), are segmented into multiple clusters using a Simple Linear Iterative Clustering (SLIC) algorithm. The skeleton is then extracted from segmented RGB and depth images via Euclidean Distance Transform (EDT). Human joints, extracted from the skeleton, provide the annotations for accurate pixel-level labeling. An elliptical human model is then generated via a Gaussian Mixture Model (GMM). A Conditional Random Field (CRF) model is trained to allocate a specific label to each pixel of different human body parts and an interaction object. Two semantic feature types that are extracted from each labeled body part of the human and labelled objects are: Fiducial points and 3D point cloud. Features descriptors are quantized using Fisher's Linear Discriminant Analysis (FLDA) and classified using K-ary Tree Hashing (KATH). In experimentation phase the recognition accuracy achieved with the Sports dataset is 92.88%, with the Sun Yat-Sen University (SYSU) 3D HOI dataset is 93.5% and with the Nanyang Technological University (NTU) RGB+D dataset it is 94.16%. The proposed system is validated via extensive experimentation and should be applicable to many computer-vision based applications such as healthcare monitoring, security systems and assisted living etc.

INDEX TERMS 3D point cloud, fiducial points, human-object interaction, pixel labeling, semantic segmentation, super-pixels, K-ary tree hashing.

I. INTRODUCTION

Understanding Human-Object Interaction (HOI) is formulated on Human Action Recognition (HAR) [1]. However, HOI is not limited to identify human actions, it can also detect relationships between humans and objects [2]. This relationship is called the verb or the interaction between a human and an object. Hence HOI is called the identification of triplets (human, verb, object) [3], [4]. It is a challenging field and is of particular interest in research. Due to the complex nature of HOI there is a need for a very thorough understanding of each movement involved in an interaction. Semantic segmentation has proved to be very effective in multiple domains of image processing and computer vision, such as intelligent transportation, medical imagery, object detection and human-computer interaction [5], [6]. Semantic segmentation is the clustering of pixels that belong to the same class and labeling them individually [7]. Therefore,

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

we semantically segment different human body parts and their interaction object. Traditional HOI recognition systems based on semantic segmentation only consists of human and object labeling [8], [9]. However, in the proposed system, different human body parts along with their respective object are semantically segmented and labelled. In this way, movements performed by each body part are recorded individually resulting in the development of an accurate HOI recognition system.

Human-object interaction is a very popular research area due to its wide applicability to, e.g., healthcare monitoring [9], assisted living [10], surveillance [11], motion sensing games [12], content-based video indexing and retrieval, etc., in the field of computer vision [13], [14]. Thus, there is a need for a more reliable and accurate system. A lot of research work has been performed in recent years in this field. Nevertheless, there remain some challenges that need to be tackled, such as variation in lighting and occlusion of different objects [15], [16]. In order to overcome these challenges, we propose a fusion of RGB and depth sensors. Depth sensors overcome the problem of occlusion and prove to be very effective in action recognition by providing the extra depth information of each object involved in an interaction [17]. Moreover, we incorporate semantics in each action class which results in deeper understanding of each movement performed by each body part during interaction.

The proposed HOI recognition system consists of four major modules: image normalization, human and object segmentation, human body parts and object detection via elliptical modeling and pixel-level labeling, HOI interaction recognition via semantic feature extraction, dimensionality reduction and classification. First, RGB and depth images are normalized by removing noise via Bilateral Filtering (BLF). The de-noised images are subjected to a segmentation phase performed via a Simple Linear Iterative Clustering (SLIC) algorithm. After the segmentation of both humans and objects from the backgrounds, skeletons are extracted via Euclidean Distance Transform (EDT) to trace the human skeleton human. The branch points of skeletons are given as centroids to form clusters of different body regions by Gaussian Mixture Model (GMM). The orientation and region under each cluster is represented by an ellipse. Hence an elliptical model representing different body parts and their respective objects is produced.

After modeling each human and object, pixel-level labeling of each region under an ellipse is performed. Conditional Random Field (CRF) is trained to label each RGB and depth image. Two unique features from each labelled human body part and object are extracted. After feature extraction, Fisher's Linear Discriminant Analysis (FLDA) is used for the dimensionality reduction of feature descriptors. In the end, each action class is classified and recognized via a K-ary Tree Hashing (KATH) classifier. The efficiency of the proposed work is validated via experimentation over three datasets: the Sports dataset, the Sun Yat-sen University (SYSU) 3D HOI dataset and the Nanyang Technological University (NTU) RGB+D dataset.

The major contributions of this work can be summarized as follows.

- Improved silhouette segmentation for both RGB and depth images via a SLIC algorithm.
- A precise human body parts detection and ellipsoidal model that is generated from detected human body parts.
- Pixel-level labeling of each detected human body parts and object from both RGB and depth image sequences via CRF.
- The main contribution is accurate HOI detection via unique semantic feature extraction, from each labeled body part and object.

The rest of the paper is structured as follows: Section II provides the related work. Section III presents the detail of each module of the proposed HOI system. Section IV describes the experiments performed for validating the performance of the system and comparison of the recognition rate of our work with other systems. Finally, Section V provides the conclusion with some future directions.

II. RELATED WORK

Many HOI recognition systems have been proposed in recent years comprising of both deep learning [18]–[20] and machine learning based approaches [21]. However, in our proposed work, we have developed a machine learning based multi-vision sensors system that incorporates a semantic segmentation technique. Therefore, we divide the related work into two sections. The first section describes related work that reports multi-vision sensors based HOI. The second section consists of action recognition systems based on different semantic segmentation techniques.

A. MULTI-VISION SENSORS BASED HOI SYSTEMS

Data acquisition in vision-sensors based action recognition systems comprise of RGB [22], [23], depth and skeletal data [24], [25]. In this section related work in the field of HOI systems based on all three aforementioned vision sensors techniques is presented. Yao and Fei-Fei [26] proposed an HOI system that consists of a mutual context for human and object. The two types of contextual data used in this method are: co-occurrence context models and the co-occurrence statistics between objects and human poses. Furthermore, to represent the relationships between humans and objects, a spatial context is also represented. The efficacy of the system is proved with two publically available RGB datasets but still, the system lacked annotation of human body parts and objects. Yan *et al.* [27] proposed a multitask neural network based HOI recognition system based on a combination of human body and hand motion. A digital glove called WiseGlove was used to record the motion of the hands. A neural network based technique was used to identify object and HOI. Experimental results with both RGB and skeletal data achieved a better recognition rate but testing was performed with a very limited data range of eight action classes.

Meng *et al.* [28] proposed a system based on the distance of skeletal joints. This is a depth sensor-based system in which inter-joint and joint-object distance is calculated. The features in this system were pose invariant and classified by random forest. A good recognition rate was achieved but the system was tested on only one dataset. In [29] Wan *et al.* proposed a pose aware system for HOI detection. A global spatial configuration of HOI is captured to focus only on action related parts of humans. In order to incorporate pose, a multi-branch network is used to represent the relationships between semantic parts, objects and interaction contexts. Two publically available RGB datasets were used for experimentation. Robust predictions were made via fine grained HOI recognition rates. Qi *et al.* [30] proposed Graph Parsing Neural Network (GPNN) based HOI detection. The graph structure includes the adjacency matrix and node labels. Results on two RGB and one depth dataset proved the validity of the system. In [31] Gkioxari *et al.* proposed the detection of triplets, i.e., human, verb and object detection. They exploit the concept of the appearance of a person to determine the object and the interaction. An action-specific density is calculated to detect the targeted object. They proved the effectiveness of their approach through extensive experimentation on two RGB datasets. Li *et al.* [32] proposed a 3D pose based system and a new benchmark named Ambiguous-HOI. To mine features, a 2D and 3D representation network is proposed. To represent both humans and objects, a cross-modal consistency tasks and joint learning structure was proposed. Experimentation on two RGB datasets proved the better performance of the system.

B. SEMENTIC SEGMENTATION AND LABELING

Semantic segmentation is the significant part of our proposed methodology. In recent years, different methodologies have been adopted by researchers in the field of semantic segmentation [33], [34]. So, in this section we describe some semantic segmentation based scene understanding and human action identification systems. In [35], Zhou *et al.* proposed cascaded parsing network based HOI. An instance detection module and interaction reasoning module were proposed. HOI representation, in the form of instance and relation features, is parsed via GPNN. The detection of interaction is not only limited to a bounding box but to pixel-level segmentation of humans and objects. Experimental results demonstrate better performance than prior methods on two RGB datasets (V-COCO and HICO-DET). In [36], J. Ji *et al.* provided semantic segmentation for different action classes instantaneously. They used the concepts of multi-task learning and contextual data. Region Based Convolutional Neural Networks (R-CNN) was used for pixel-level labeling. They proved the performance of the system by experimenting both detection and segmentation on one RGB dataset. Khowaja and Lee [37] proposed a semantic analysis of videos by applying localized sparse segmentation using global clustering. Through experimentation they proved that semantic images produce better activations by focusing on regions that are significant for action recognition. The use of approximate rank pooling from Long Short Term Memory (LSTM) showed better performance. High recognition rate with three public datasets validated the performance of the system.

In [38], Arnab *et al.* proposed semantic segmentation based scene understanding via Conditional Random Field (CRF) and neural networks. They used deep neural networks to automatically learn features. The mean field algorithm of CRF was used as a Recurrent Neural Network (RNN) layer. They improved the segmentation performance on an RGB public dataset (Pascal VOC). Paisitkriangkrai *et al.* [39] proposed a pixel labeling technique consisting of CRF and Convolutional Neural Networks (CNN). They proposed robust features by combining both hand-crafted and CNN extracted features. Then, to label probabilities, CRF was applied. As a result, segmentation was improved with the ISPRS labelling contest dataset. In [40] A. Jalal proposed a depth silhouette based HAR labeled human body parts and identified the centroids of each part. Random field was used to label and train the images. A motion vector comprising of magnitude and direction was computed via identified centroids. Experiments performed on six daily life activities show a better recognition rate than many state-of the art methods. All of these methodologies showed improvement in the recognition of human actions so we propose an HOI recognition system based on semantic human body parts segmentation.

III. THE PROPOSED APPROACH

The proposed approach consists of four major modules: image normalization, human and object segmentation, human and object detection, modeling and labeling and in the end HOI recognition. The overall architecture of the proposed system is shown in Fig.1. Detail of the techniques used for each of the aforementioned modules is explained in the following subsections.

A. IMAGE NORMALIZATION

During pre-processing, the raw images from both RGB and depth datasets are fed into the system. In order to keep dimension of images from all three datasets similar, they are cropped to a fixed dimension of 560×350. After identifying the initial region of interest, BLF is applied. BLF removes noise, smooths the images and preserve the edges of all the objects in the images [41]. All the images are de noised by Gaussian smoothing kernels. The intensity value of each pixel x of the image I is replaced by a weighted intensity obtained by neighboring pixels [42]. The range kernel f_r smooths differences in intensities and spatial kernel g_s smooths differences in coordinates. The filtered image I^{fil}, obtained after applying bilateral filter, is defined as;

$$I^{fil}(x) = \frac{1}{W_p} \sum_{x_i \in \Omega} I(x_i) f_r(\|I(x_{i)} - I(x)\|) g_s(\|x_i - x\|) \quad (1)$$

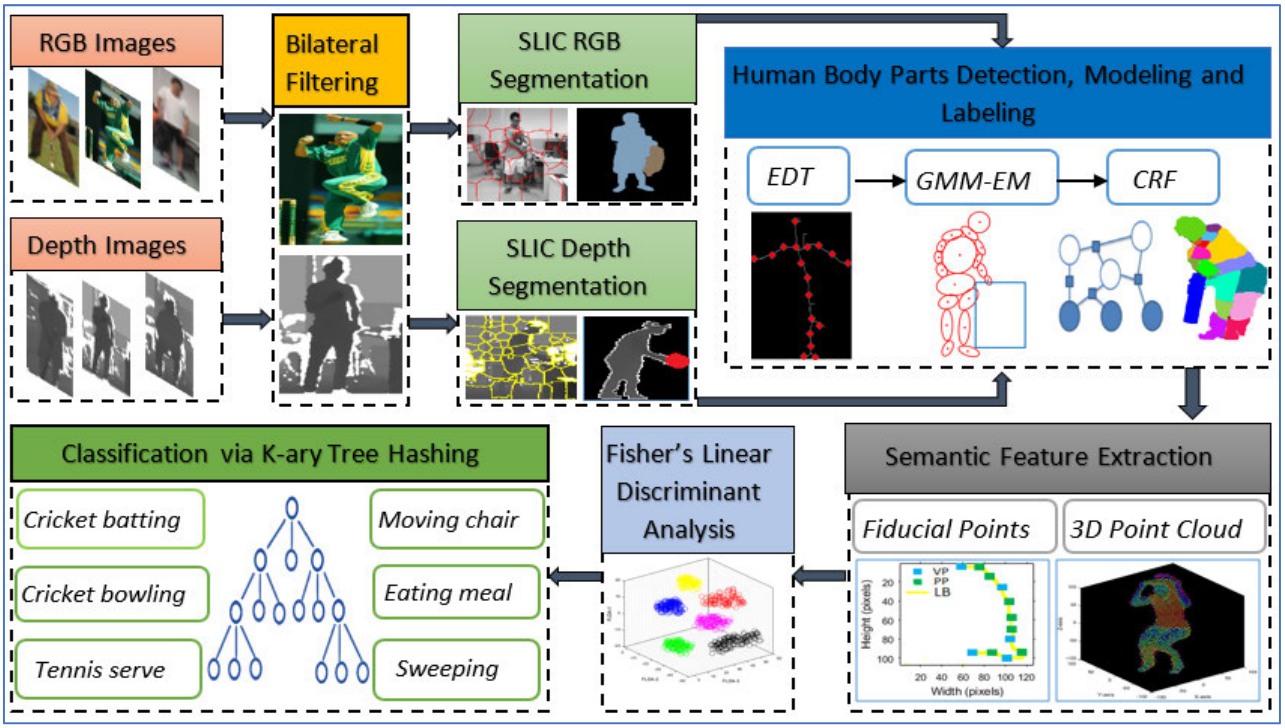


FIGURE 1. Architecture of the proposed HOI recognition system.

where x_i is one of the neighboring pixels from the specified neighborhood window Ω centered at x and W_p is defined as;

$$W_p = \sum_{x_i \in \Omega} I(x_i) f_r(\|I(x_i) - I(x)\|) g_s(\|x_i - x\|) \quad (2)$$

The weight of W_p, i.e., the normalization factor is assigned by spatial closeness and differences in intensity values.

B. HUMAN AND OBJECT SEGMENTATION

After the normalization phase, both humans and their respective interaction objects are segmented from the background using an SLIC algorithm [43]. In this section, we propose a linear iterative clustering based super-pixels approach. In this approach k- means the algorithm is used to generate super-pixels [44]. First, all RGB images are converted to a lab (l* specifies lightness, and a* and b* for the four colors: red, green, blue, and yellow) color space. After that, the numbers of super-pixels k is specified. Then, initial centers $C_i = [l_i\, a_i\, b_i\, x_i\, y_i]^T$, which are S pixels apart, are initialized for each cluster. The grid interval $S = \sqrt{N/k}$ produces nearly equal sized super-pixels. The centers of the super-pixels should not be at the edges of objects, for this reason the centers are moved to the lowest gradient positions in a 3×3 neighbourhood. After specifying a cluster, searching starts where each pixel i is assigned to its nearest cluster center. Compared to traditional k-means, the search space of the SLIC algorithm is very limited [45]. The search space is reduced by measuring the distance D to define the nearest centers for each pixel. This distance D is a 5D Euclidean distance in a labxy color space and is given by 3D color distance d_c and 2D spatial distance d_s as;

$$d_c = \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2} \quad (3)$$

$$d_s = \sqrt{(x_j - x_i)^2 + (y_j - y)^2} \quad (4)$$

As the depth images are given in grayscale, they only have l component so the distance d_c for grayscale is given as $d_c = \sqrt{(l_j - l_i)^2}$.The normalized distance D' is given by maximum distance within the cluster in color and space proximity N_c and N_s respectively as;

$$\mathrm{D}' = \sqrt{\left(\frac{d_c}{N_c}\right) + \left(\frac{d_s}{N_s}\right)} \quad (5)$$

At the final stage, after each pixel is assigned to the nearest neighbor, an update in cluster centers is made. At the end there are still some pixels that are not assigned to their respective clusters. So, a super-pixel merging algorithm [46] is performed to further refine the segmentation process. Different visual features are used to define n super-pixels $X = [x_1, \ldots, x_n] \in R^{m \times n}$ in an image. These image features describe l semantic labels in an image and the similarity of any super-pixel x_i and x_j is given as;

$$S_{i,j} = \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \left[\delta_1 d_{ij}^{lab} + \delta_2 d_{ij}^{tex} + \delta_3 d_{ij}^{sift} + \delta_4 d_{ij}^{surf}\right] \times D_{i,j} \quad (6)$$

where δ is the weight factor for distance adjustment, $d_{ij}^{lab}, d_{ij}^{tex}, d_{ij}^{sift}, d_{ij}^{surf}$ represent the Euclidean distance between color, texture, sift and surf distances of super-pixels

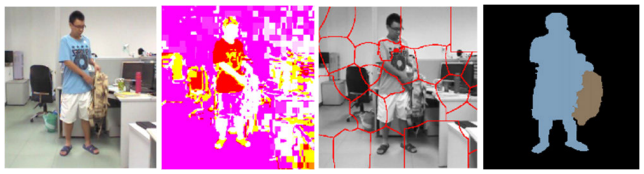


FIGURE 2. Results of the SLIC algorithm on RGB images showing (a) the original RGB image, (b) converted L*a*b* color space, (c) super-pixels overlaid on the grayscale image and (d) segmented regions after super-pixel merging.

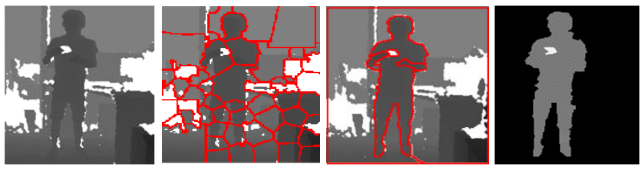


FIGURE 3. Results of the SLIC algorithm on depth images of phone interaction showing (a) original images, (b) super-pixel extraction, (c) super-pixel merging and (d) the segmented human and object.

i and j. The relationship between super-pixels is stored in D. If c_i is adjacent to c_j then $D_{i,j} = 1$, $D_{i,j} = 0$ otherwise. Fig. 2 shows the results of SLIC segmentation over an RGB image while Fig.3. shows the results of the SLIC algorithm on depth images from the SYSU dataset.

C. HUMAN BODY PARTS DETECTION, MODELING AND LABELING

In this section human body parts are detected, elliptically modeled and labeled. This section is divided into three phases. In the first phase, human body parts are detected and in the second phase an ellipsoidal model of human body parts is generated. In the third phase, detected human body parts and the respective interaction object are labelled using CRF. Each of these phases is described in the following sub-sections.

1) HUMAN BODY PARTS DETECTION

In order to provide accurate human joints annotations, a human skeleton is first extracted via Euclidean Distance Transform (EDT) [47]. All segmented RGB and depth images are converted to binary images. The binary images are then converted to grayscale images in which only those foreground pixels p are taken whose distance from the background pixels q is minimum [48]. This grayscale image is called the Distance Transform [DT] and its pixel values are given by:

$$DT_p = \min\{d\,(p.q)\,|I\,(q) = 0\} \tag{7}$$

This distance is calculated by Euclidean distance. Then a morphological operation of thinning is the applied to extract continuous skeleton pixels [49]. The operation of skeletonization further reduces the image to a single line without destruction of the structure of a human. The skeletonization process is demonstrated in Fig. 4.

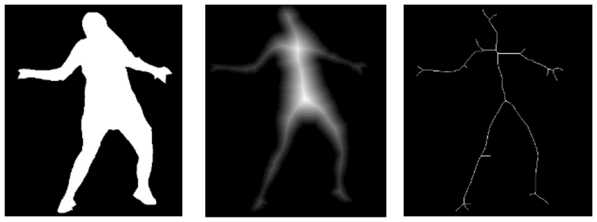


FIGURE 4. Skeletonization results showing, (a) the original binary images, (b) EDT and (c) the extracted skeleton via morphological thinning.

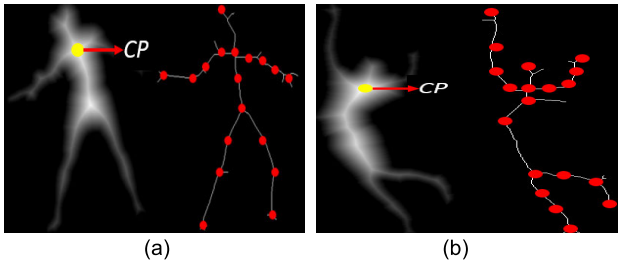


FIGURE 5. Skeletal joints detection via branch points extraction on (a) tennis forehand and (b) volleyball smash interactions.

In order to determine skeletal joint points, first, a Critical Point (CP) is determined [50]. This is the point with a minimal *DT* from the boundary. Keeping the CP as the root node, a tree is traversed in four directions, i.e., upward, downward, left and right. In each of these directions, a constraint is followed, i.e., only the foreground pixels having value of 1 are searched in 8-connected neighborhood [51]. There are two types of points in this search.

- *Endpoints(EP)*: those points in which there is only one skeletal point among 8 neighborhoods.
- *Bifurcation points(BP)*: those points in which there are three or more than three skeletal points among 8 neighborhoods.

By keeping the root node as a *parent node*, the first *child* is searched in the upward direction. The direction of search is guided by the slop of the line that is connecting the root. The EP in the upward direction is the head and the BP is the neck. In the search from root node to the left direction, the EP is the left hand. The mean of the EP and the CP is the left elbow and the mean of the left elbow and the neck is the left shoulder. Similarly, these three joints of right arm are located by a search towards the right. The root or the CP is the torso of the human body. Searching from the CP in the downward direction, the first BF is the torso base. The EP in the left side is the left foot and, on the right side, it is the right foot. The mean point between the torso base and both feet is the left knee and the right knee point respectively. Similarly, the mean point between the torso base and both knees is the left and right upper legs, and the mean point between both the knees and both the feet is the right and left ankle joint respectively. In this way a total of eighteen human joints are located as shown in Fig. 5. During all this searching, the constraint of the pixel value as 1 is followed.

2) GAUSSIAN-BASED ELLIPTICAL MODELING

The binary image I with skeleton S and k numbers (i.e. 18) of joint annotations is fed to an elliptical modeling phase. As the object is already detected in section (B) so in this section a Gaussian Mixture Model-Expectation Maximization (GMM-EM) algorithm [52] is implemented to represent each body part with an ellipse. First of all, the human is represented by non-overlapping or partially overlapping circles CC. A graph $G = (V, W)$ in which V is a set of the nodes that represent the endpoints of a skeleton and W is the set of edges. The skeleton is segmented into l_i parts by W where $i \in \{1, \ldots, |W|\}$ and W is the number of edges. From these segmented parts the 16-bin histogram of radii of each circle and shape complexity C is computed where C is an entropy function given as;

$$C = -\sum_{i=1}^{|W|} \sum_{j=1}^{16} p_{ij}\, log p_{ij} + log\,|S| \tag{8}$$

Through these circles, ellipsoidal fitting within the boundary is initiated. Each joint location is the centroid c of the circle and the line joining two consecutive joints in the skeleton is the radius R of each circle. Each of these circles is tangent to the boundary at two points. The ellipse fitting process is carried out according to the GMM-EM algorithm. The GMM-EM algorithm is initialized by specifying the centroid and the number of clusters because random initialization will have led to a suboptimal local minimum of the problem [53]. The shape skeleton by EDT is exploited for a more informed decision at the GMM-EM initialization stage. The 2D Gaussian function used for clustering of foreground pixels is given by;

$$P_i\,(\mathrm{p}) = A_i.e^{-(p-c_i)^T M_I (p-c_i)} \tag{9}$$

where P_i gives the probability of a foreground pixel p to belong to an ellipse E_i with origin c_i and M_i represents a 2×2 matrix with eccentricity ad orientation information of E_i. Moreover, the amplitude $A_i = 1$ to keep the same values of $P_i\,(\mathrm{p})$ at the ellipse boundary for all ellipses. In this way the probability of a particular point belonging to an ellipse depends only on its position, orientation and eccentricity and not on the area of the ellipse. The object is already detected in Section B using SLIC. In this section the object is enclosed with a bounding box using connected components and blob analysis on the segmented image. The elliptical models for some sample images from SYSU and NTU datasets are shown in Fig. 6.

3) PIXEL-LEVEL LABELING OF HUMAN BODY PARTS AND THEIR RESPECTIVE OBJECTS VIA CRF

After detecting different human body parts, the results of the elliptical modeling phase are fed to the pixel-level labeling phase, as each human body part is already segmented with an ellipse in the previous phase. In this phase, fully-connected CRF [54] is used to assign a label for each pixel in each of the detected human body parts and the object. A relationship between the output variables (specified labels) represented as $y = y_1, y_2, \ldots . y_N$ and observed features (such as pixel intensities) as input variable x is described by CRF in the form of conditional probability $P\,(y\,|\,x)$ [55]. During pixel labeling, N is the total number of pixels and y_i is the label assigned to the i^{th} pixel. The modeling of $P\,(y\,|\,x)$ in CRF is approached by representing y as a Markov random field. CRF is represented in the form of an undirected graph as $G = (V, E)$ with V as a set of vertices or nodes and E as a set of edges of the graph. Each label y_i corresponds to each node [56]. In order to assign a label to each pixel, the probability distribution of CRF is defined in the form of an energy function as;

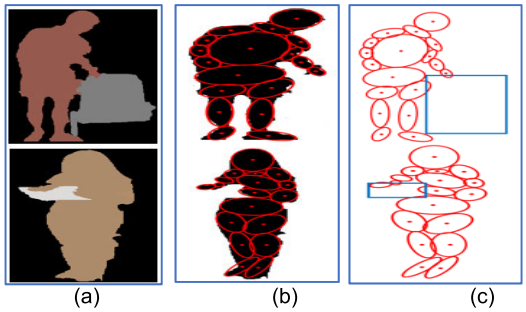


FIGURE 6. Elliptical models showing (a) the original segmented image, (b) the ellipse fitting of human body parts with joint annotations and (c) the final human body parts and object model.

$$P\,(y\,|\,x) = \frac{1}{Z_x} exp\,\{-E\,(y; x)\} \tag{10}$$

where Z_x is the partition function to normalize the probability distribution and $E\,(y; x)$ is the energy function that is the sum of smaller clique potentials $\psi_c(y_c;\,)$ represented as;

$$E\,(y; x) = \sum_c \psi_c(y_c; x) \tag{11}$$

So, to define a label for each pixel, an energy function is defined as;

$$E\,(y; x) = \sum_{i=1}^{n} \psi_i^U\,(y_i; x) + \sum_{\bar{y} \in \varepsilon} \psi_{ij}^P(y_i, y_j; x) \tag{12}$$

where ψ^U is the unary energy component associated with each pixel and ψ^P is the pairwise energy component associated with a set of pixels ε. The most probable assignment to a label requires minimization of the energy function as;

$$\hat{y} = \underset{y}{\mathrm{argmax}}\, E(y; x) \tag{13}$$

For inference, a mean-field algorithm is used as given in Algorithm 1 that approximates energy minimization [57]. This algorithm is initialized with Gibbs distribution and it performs in a loop for Q energy minimization. The weighted Gaussian is computed in a message passing step.

In order to train the data, maximum likelihood is used. The parameters that produce the training data with the highest probability under the model are chosen. On a training sample $((x_1, y_1)\,, (x_2, y_2) \ldots (x_T, y_T))$ conditional log likelihood is

Algorithm 1 Mean Field Inference Algorithm

1. Initialize Q as $Q_i(l) \leftarrow \frac{1}{Z_i} exp\{-U_i(l)\}$ for all i pixels
   while not converged **do**
2. Message passing from all X_j to all X_i as;

$$Q_i^{(m)}(l) \leftarrow \sum_{j \neq i} k^{(m)}\left(f_i, f_j\right) Q_J\left(l\right) \tag{14}$$

   //where $k^{(m)}\left(f_i, f_j\right) Q_J$ is Gaussian kernel and $\left(f_i, f_j\right)$ is feature vector for pixel i and j//
3. Adding weight to filter outputs as $Q_i^{(l)} \leftarrow \sum_m w^{(m)} Q_i\left(l\right)$ for all m
   //where $w^{(m)}$ is the weight of m-th kernel//
4. Compatibility transform

$$Q_i^{(l)}(l) \leftarrow \sum_{l' \in L} \mu\left(l, l'\right) Q_i\left(l'\right) \tag{15}$$

5. Adding unary potentials to $Q_i^{(l)}(l)$ as $Q_i^{(l)} \leftarrow U_I\left(l\right) - Q_i(l)$
6. At the end normalizing output as $Q_i^{(m)} \leftarrow \frac{1}{Z_i} exp\left(Q_i(l)\right)$

end while

maximized with respect to unknown parameter θ as;

$$\begin{aligned}&(\arg)\max_{\theta} \sum_{t=1}^{T} \ln p\left(y_t \mid x_t; \theta\right) \\ &= (\arg)\min_{\theta} \sum_{t=1}^{T} \left[-lnZ\left(x_t, \theta\right) - E(x_t, y_t, \theta)\right]\end{aligned} \tag{16}$$

Now a CRF model is trained to predict a correct label for each segmented body part and the interaction object. Some of the results of the CRF in the form of labelled body parts and objects are demonstrated in Fig. 7.

D. HUMAN-OBJECT INTERACTION RECOGNITION

HOI recognition is the identification of triplets (human, verb and object) as the human along with its body parts and interaction object are detected and labelled. Now, in this section, the verb, i.e., the interaction between the human and the object is identified. For HOI interaction recognition, this section is sub-divided into three modules. The first is the semantic feature extraction, the second is dimensionality reduction via FLDA, and the third phase is classification with KATH.

1) SEMANTIC FEATURE EXTRACTION

The two types of features extracted from each semantic region including human body parts and objects are fiducial points and 3D point cloud.

a: FIDUCIAL POINTS

The Fiducial Point (FP) of each human body part and each object is detected individually [58]. First of all, the boundary of each segmented body part is detected and then divided into Left Boundaries (LB) and Right Boundaries (RB). The boundary points are scanned from top to bottom in a horizontal rows of the xy coordinates. In a i^{th} row the transition in x-axis from high pixel values to low pixel values indicates right boundaries as $RB = \{rb_1, rb_2, \ldots . rb_m\}$. Similarly, the transition along the x-axis from low pixel values to high

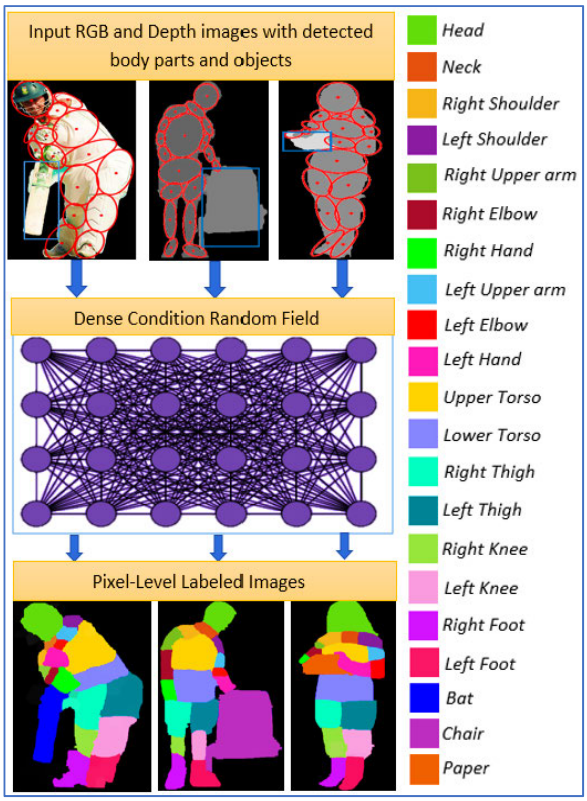


FIGURE 7. Pixel-level labelling over samples of (a) the sports dataset, (b) the SYSU dataset and (c) the NTU RGB+D dataset.

pixel values indicates $LB = \{lb_1, lb_2, \ldots . lb_n\}$. Where m and n are the total numbers of pixels in RB and LB respectively. After identifying RB and LB, peaks and valley points are detected in each side of the boundary. The first-order derivative is taken as local maxima and minima to detect peaks and valleys respectively. A change in the slope of a boundary from negative to positive is referred to as minima while a change in slope from positive to negative is referred as maxima. The first order derivative of RB from $i = 1 \ldots . m - 1$ is $drb_i = rb_{i+1} - rb_i$ and its vector is given as;

$$dRB = drb_1, drb_2, \ldots \ldots drb_{m-1} \tag{17}$$

The first order derivative of LB from $j = 1 \ldots . n - 1$ is $dlb_j = lb_{j+1} - lb_j$ and its vector is given as;

$$dLB = dlb_1, dlb_2, \ldots \ldots dlb_{n-1} \tag{18}$$

The peaks P_r of RB are given as;

$$P_r = \{rb_i \mid drb_i \geq 0 \cap drb_{i+1} < 0\}, \quad \forall_i = 1, 2 \ldots m - 1 \tag{19}$$

Similarly, the peaks P_l of LB are calculated from dLB. On the other hand the valleys V_r of RB are calculated as;

$$V_r = \{rb_i \mid drb_i \leq 0 \cap drb_{i+1} > 0\}, \quad \forall_i = 1, 2 \ldots m - 1 \tag{20}$$

Similarly, the valleys V_l of LB are calculated from dLB. If the contour of any body part is flat, i.e., if it has consecutive

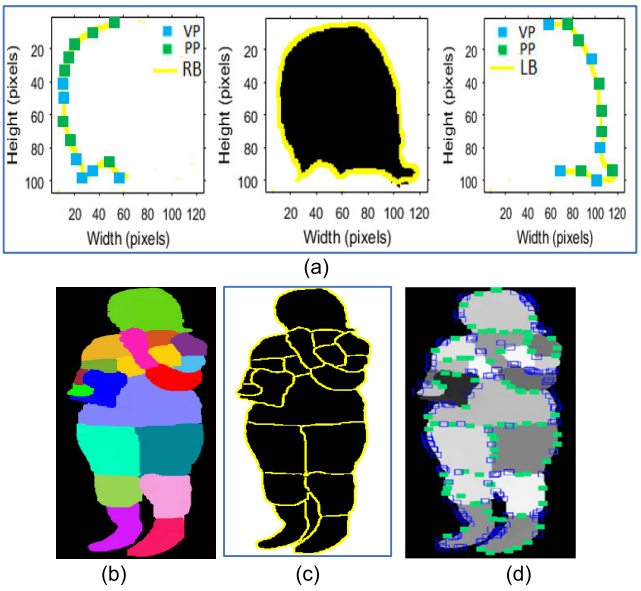


FIGURE 8. Peaks and valley points detection showing (a) PP and VP detection on boundaries of the head, (b) labeled eating meal interaction, (c) detected boundaries on each body part and (d) detected PP and VP on each body part.

TABLE 1. Fixed numbers of valley points and peaks for each human body part and object.

Human Joints	VP No.	PP No	Human Joints	VP No.	PP No.
Head	15	15	Upper Torso	60	60
Neck	15	15	Lower Torso	60	60
Right Shoulder	20	20	Right Thigh	30	30
Left Shoulder	20	20	Left Thigh	30	30
Right Upper arm	15	15	Right Knee	15	15
Left Upper arm	15	15	Left Knee	15	15
Right Elbow	15	15	Right Foot	15	15
Left Elbow	15	15	Left Foot	15	15
Right Hand	20	20	Object	20	20
Left hand	20	20	Total VP+PP	860	

zeros, then the median point of the consecutive zeroes is taken. The coordinates value of each FP is recorded in a feature vector and tracked with each changing frame. Peaks and valley point detected on boundaries of some body parts are displayed in Fig.8.

The Peak Points (PP) and Valley Points (VP) of all the body parts and the object may not be the same in every interaction class. So these points are fixed for each body part. Table 1 shows the number of peaks and valley points for each body part and object.

b: 3D POINT CLOUD

In this feature, humans along with their interaction objects are represented in the form of point clouds. The RGB labelled images are converted into 3D point clouds with xyz coordinates [59]. Let K be a point cloud then its coordinate is given as $X_p^k = (x_k, y_k, z_k)$. The pixels in the RGB image are converted into 3D points. This conversion is made on the basis of pixel coordinates and their corresponding intensity values. In order to extract features from the point cloud, these points are down sampled using a Voxel Grid (VG) filter [60]. A voxel is a grid defined over 3D point clouds. A spatial average is taken inside each voxel to down sample the points. Those points, which lie inside the voxel bounds, are joined to form one output point. The points inside the voxel are given with the centroid as;

$$x = \frac{1}{S}\sum\nolimits_{(x,y,z)\in A} x \tag{21}$$

$$y = \frac{1}{S}\sum\nolimits_{(x,y,z)\in A} y \tag{22}$$

$$z = \frac{1}{S}\sum\nolimits_{(x,y,z)\in A} z \tag{23}$$

where S is the number of points in a voxel A. In every interaction class, each human body along with the interacted object is down sampled to 6000 cloud points while maintaining the posture or shape of the human and interaction object as shown in Fig.9.

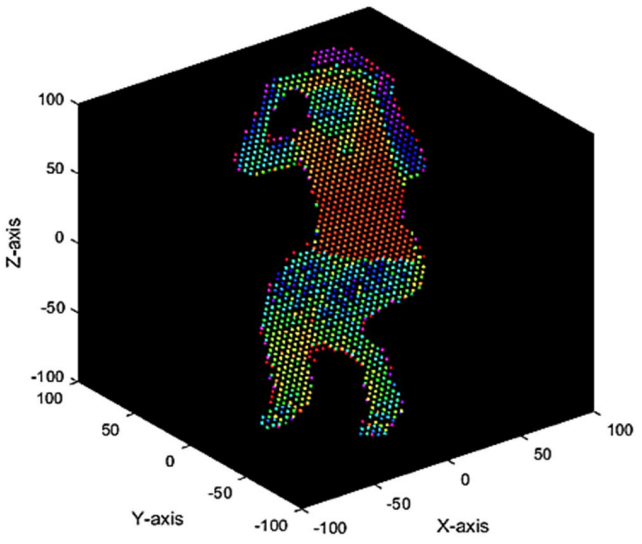


FIGURE 9. A down-sampled 3D point cloud on a human and object over a wearing hat interaction.

The coordinate value along with the intensity of each point in a down-sampled point clouds is stored in a feature descriptor. The feature descriptor from the two extracted features of both the human and object are concatenated at the end (see Algorithm (2)).

Algorithm 2 Semantic Feature Extraction

Input: Labelled images
Output: Feature Vectors containing Semantic Features

1. F1 = ExtractFiducialPoints←Human
2. F2 = Calculate3DPointCloud←Human
3. F3 = ExtractFiducialPoints←Object
4. F4= Calculate3DPointCloud←Object
5. V_i = CreateFeatureVectorsf_i
6. For each f_i in features do

{
Concatenate = (V_i, V_{i+1})
}

2) FISHER'S LINEAR DICRIMINANT ANALYSIS

After combining feature vectors of all the interaction classes, a complex matrix is generated. FLDA is used as a dimensionality reduction algorithm before classification. The objective is to reduce intra-class variance and to increase inter-class variance [23]. The FLDA is applied to the interaction classes of each dataset individually. In a multiclass discriminant analysis, let μ_i be the mean of each HOI class C, Σ be the same covariance and μ be the mean of class means, then the scatter between each class C is defined as;

$$\Sigma_b = \frac{1}{C}\sum_{I=1}^{C}(\mu_I - \mu)(\mu_I - \mu)^T \tag{24}$$

while T is the transpose and the separation of classes is given in direction $\vec{w}$ as;

$$S = \frac{\vec{w}^T \Sigma_b \vec{w}}{\vec{w}^T \Sigma \vec{w}} \tag{25}$$

The rows represent the number of images in the training set of each dataset. The final dimension of the Sports dataset after feature reduction is 6120×250, the SYSU dataset is 6120×360 and that of the NTU dataset is 6120×380. The scatter plot for the Sports and NTU datasets are displayed in Fig.10.

3) K-ARY TREE HASHING

The optimized vectors of all three classes are fed to a KATH classifier. It is a graph-based classifier given as $G = \{g_i\}$ while $i = 1 \ldots .N$ and N is the total number of objects in the graph [61]. The graph consists of vertices V, undirected edges E and label function $l : V \rightarrow L$ to assign labels to nodes in g_i from a label set L. Based on the structure of the graph and node labels, a class label y_i is also given to each graph g_i. Furthermore, a size K of the traversal table and MinHashes $\{D^{(r)}\}_{r=1}^{R}$ for R iterations is also specified [62]. Random permutation functions $\{\pi_d^{(r)}\}$ are generated for MinHashes. A MinHash technique is used to measure the Jaccard similarity $\hat{J}$ of two sets S_i and S_j based on D Minhashes as;

$$\hat{J}(S_I, S_J) = \frac{\sum_{d=1}^{D} \mathbf{1}\left(\min\left(\pi_d\left(S_i\right)\right) = \min\left(\pi_d\left(S_j\right)\right)\right)}{D} \tag{26}$$

where, if the state is true, then $1(state)$ is 1, otherwise it is 0. The KATH algorithm consist of three steps, namely, traversal table construction, recursive leaf extension and leaf sequence using the MinHash scheme. This MinHash scheme classifies the data into various interactions. All three steps are given in Algorithm 3.

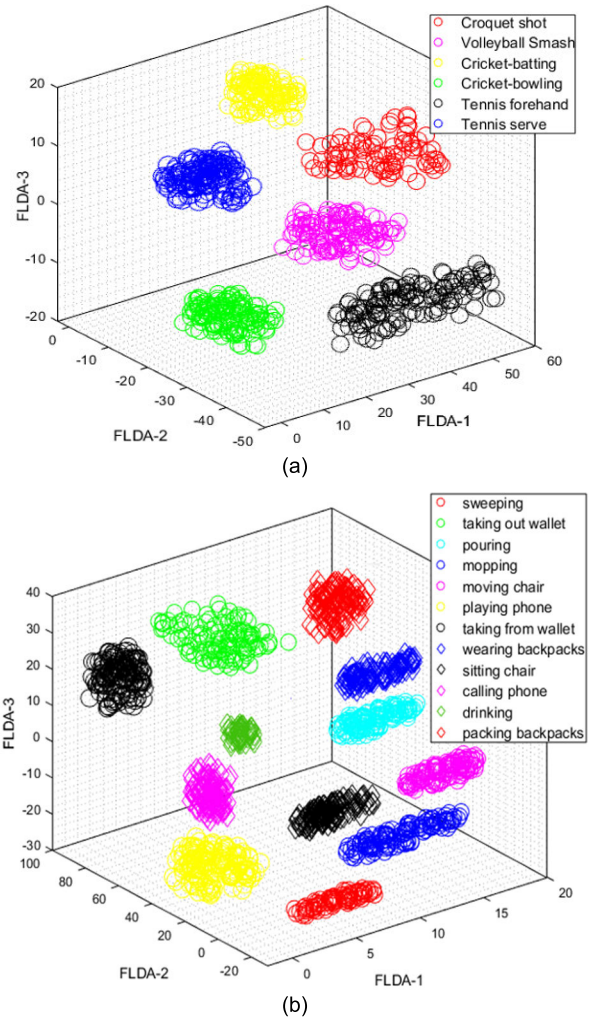


FIGURE 10. Scatter plot showing classes for (a) the sports dataset and (b) the SYSU dataset.

IV. EXPERIMENTAL SETUP AND RESULTS

This section gives the details of each experiment performed to validate the proposed system. All the processing and experimentation is performed on MATLAB (R2018a). The hardware system used is Intel Core i5 with 64-bit Windows-10. The system has an 8 GB and 5 (GHz) CPU. We divided the experiments into two sections. In the first section HOI recognition performance in which recognition accuracies of each interaction class is given in the form of a confusion matrix and the precision, sensitivity, specificity and F1 scores are also measured along with comparisons of the proposed method with other state-of-the-art (SOTA) methods. In the second section, i.e., pixel-level labeling, the label accuracies for each human body part and object are measured using CRF. All these experiments are performed using a Leave One

Algorithm 3 K-Ary Tree Hashing

Input: $g = (V, E, L)$, K, $\{D^{(r)}\}_{r=1}^{R}$
//where R is number of iterations//
Output: $\{x^{(r)}\}_{r=1}^{R}$
// Traversal Table Construction//
1. $V \leftarrow |V|$
2. $l(V+1) \leftarrow \infty$
3. $T \leftarrow (V+1) * ones(V+1, 1+K)$
for v=1: V **do**
4. $N_v \leftarrow neighbour(v)$
//MinHash Selection//
5. $temp \leftarrow [\min(\pi_1(l(N_v))), \ldots, \min(\pi_k(l(N_v)))]$
6. $T(v) \leftarrow [v, index(temp)]$
end for
//Recursive Leaf Extension//
7. $z^{(1)} \leftarrow [1:V]^T$
8. $S^{(1)} \leftarrow l(z^{(1)})$
for r=1: R **do**
if r >1 **then**
9. $z^{(r)} \leftarrow reshape\left(T\left(z^{(r-1)}, :\right), [1, *]\right)$
10. $S^{(r)} \leftarrow reshape\left(l\left(z^{(r)}\right), [V, *]\right)$
end if
//Leaf Sequence//
11. $f^{(r)} \leftarrow \left[h\left(S^{(r)}(1, :)\right), \ldots h\left(S^{(r)}(V, :)\right)\right]^T$
12. $x^{(r)} \leftarrow \left[min\left(\pi_1^{(r)}(f^{(r)})\right), \ldots min\left(\pi_{D^{(r)}}^{(r)}(f^{(r)})\right)\right]^T$
end for

Subject Out (LOSO) cross-validation scheme. Each dataset is divided into N subsets containing k number of images. First all the subsets are used to train the system and then one subset is used for testing. The system is then validated by taking another subset for testing and the remaining subsets for training. The images of the subsets that are used for training are not included in the testing set. In case of the sports dataset, there is a different subject in each image of each interaction class so LOSO cannot be applied. The sports dataset is divided by splitting 50% images of each interaction class for training and the rest of 50% for testing of the system. In case of the SYSU and the NTU RGB+D dataset, LOSO is applied in which the actions performed by one subject are used for testing and the actions performed by the rest of the subjects are used for training. This section is further divided into two sections: dataset description and experimental results.

A. DATASETS DESCRIPTION

The three datasets that are used for experimentation are: The Sports dataset, the SYSU 3D HOI dataset and the NTU RGB+D dataset. Details of each dataset are given in following subsections:

1) THE SPORTS DATASET

This is a static image dataset that consists of six RGB sports activities. The activities performed in this dataset are: cricket batting, cricket bowling, croquet shot, tennis forehand, tennis serve and volleyball smash. The details of the dataset and samples are given in [63]. This is a complex dataset as the poses and scenes of many interaction classes are similar to each other, e.g., volleyball smash and tennis serve. Few sample images of sports dataset are displayed in Fig. 11.

FIGURE 11. RGB samples for six interaction classes of sports dataset.

FIGURE 12. RGB and depth images of (a) mopping, (b) moving chair and (c) pouring interactions from SYSU dataset.

2) THE SYSU 3D HOI DATASET

This is an RGB-D dataset in which a Kinect sensor is used to collect RGB and depth images. This dataset consists of twelve human object interactions performed by 40 participants. The HOI classes in this dataset are sweeping, mopping, taking from wallet, taking out wallet, moving chair, sitting chair, packing backpacks, wearing backpacks, playing phone, calling phone, pouring and drinking. There are 480 video clips of different durations ranging from 1.9s to 21s. The details of the dataset and samples are given in [64]. Few RGB and depth sample images of SYSU dataset are displayed in Fig. 12.

3) THE NTU RGB+D DATASET

This is an RGB-D dataset that contains RGB, depth and 3D skeletal data. This dataset contains 56,880 video samples of 60 action classes. Only RGB data is in the form of video while the depth data is provided in the form of

image sequences. Out of 56,880 video samples provided for 60 action classes, 2715 video samples are used in the proposed work. The actions in this dataset have three categories: 40 daily actions (e.g., reading, drinking, eating), nine medical conditions (e.g., falling down, sneezing, staggering,), and 11 mutual actions (e.g., hugging, punching, kicking). From the 40 daily actions, we only worked on twelve human-object interactions. The twelve interactions that we used for training and testing in the proposed system are: drink water, eat meal, brush teeth, brush hair, tear up paper, put on jacket, take off jacket, put on a hat/cap, take off a hat/cap, phone call, play with phone/tablet and taking a selfie. The objects in these interactions are: glass of water, meal, paper, jacket, hat and phone. The rest of the details and samples are given in [65]. Few sample images of NTU REGB+D dataset are displayed in Fig. 13.

FIGURE 13. RGB and depth images showing few samples from interaction classes of NTU RGB+D dataset.

B. PERFORMANCE METRICES AND RESULTS

The two types of experiments performed for system's validation were, HOI recognition performance via KATH and pixel-level labeling performance via CRF. The results for each experiment are given in the following sub-sections.

1) HOI RECOGNITION PERFORMANCE

In this section, the performance of the system validated from a mean accuracy, precision, sensitivity, specificity and F1 scores. A comparison of the proposed system with other SOTA methods is also given in this section. The results for each performance metric is given in the following subsection:

a: HOI CLASSIFICATION ACCURACY

This experiment was repeated three times on the testing sets for each dataset individually to evaluate the classification accuracy using the KATH classifier. The results of this experiment are given in the form of confusion matrix showing true positive, true negative, false positive and false negative for each class individually. The confusion matrix for the Sports, SYSY 3D HOI and NTU RGB+D datasets are given in Tables 2, 3 and 4 respectively. It can be observed, from Tables 2, 3 and 4, that classes of all three datasets achieved high recognition rates with the mean accuracy rates of 92.88%, 93.5% and 94.16% with the Sports, SYSU and NTU datasets respectively. However, there is still some confusion between interaction classes that involve similar actions such as the tennis forehand and the volleyball smash interactions in sports dataset. Similarly, weeping and mopping interactions of the SYSU dataset are confused with each other. It can also be observed from the results of this experiment that confusion happens among the interaction classes that involve similar objects. For example, moving chair, siting chair interactions of the SYSU dataset and the phone call, play with phone and taking selfie interactions of the NTU dataset.

TABLE 2. Confusion matrix of individual HOI class over the sports dataset using KATH.

HOI Classes	Cbat	Cbow	Cro	TF	TS	VB
Cbat	**0.96**	0.00	0.04	0.00	0.00	0.00
Cbow	0.00	**0.93**	0.00	0.02	0.00	0.05
Cro	0.04	0.00	**0.95**	0.00	0.00	0.01
TF	0.00	0.00	0.00	**0.90**	0.03	0.07
TS	0.00	0.00	0.00	0.04	**0.92**	0.04
VB	0.00	0.02	0.00	0.07	0.00	**0.91**
Mean recognition accuracy = 92.88%						

*Cbat = Cricket batting; Cbow = Cricket bowling; Cro = Croquet shot; TF = Tennis forehand; TS =Tennis serve; VB = Volleyball smash.

b: PRECISION, SENSITIVITY, SPECIFICTY AND F1 MEASURES

In this experiment precision, sensitivity, specificity and F1 scores of all the interaction classes for each dataset are calculated as;

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{27}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{28}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{29}$$

$$F1\ score = \frac{2(Precision \times Recall)}{(Precision + Recall)} \tag{30}$$

The precision, sensitivity, specificity and F1 scores of the Sports, SYSU and NTU RGB+D datasets are given in Table 5, Table 6 and Table 7 respectively. It is observed from Table 5 that the positive predicted values, i.e., precision is very high for all the classes of the Sports dataset. The lowest precision of 84% is achieved with volleyball smash due to its high false positive rate. The volleyball smash also has the lowest sensitivity and F1 score due to its confusion with

TABLE 3. Confusion matrix of individual HOI class over the SYSU dataset using KATH.

HOI Classes	sweeping	mopping	TFW	TOW	Moving chair	Sitting chair	PB	WB	Playing phone	Calling phone	pouring	drinking
sweeping	**0.94**	0.05	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mopping	0.05	**0.95**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TFW	0.00	0.00	**0.93**	0.02	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00
TOW	0.00	0.00	0.04	**0.91**	0.00	0.00	0.00	0.02	0.00	0.00	0.03	0.00
Moving chair	0.00	0.00	0.00	0.00	**0.98**	0.02	0.00	0.00	0.00	0.00	0.00	0.00
Sitting chair	0.00	0.00	0.00	0.00	0.04	**0.96**	0.00	0.00	0.00	0.00	0.00	0.00
PB	0.00	0.00	0.01	0.00	0.00	0.00	**0.94**	0.04	0.01	0.00	0.00	0.00
WB	0.00	0.00	0.00	0.00	0.00	0.00	0.04	**0.95**	0.00	0.00	0.01	0.00
Playing phone	0.00	0.00	0.02	0.00	0.00	0.00	0.03	0.00	**0.89**	0.05	0.01	0.00
Calling phone	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	**0.94**	0.00	0.02
pouring	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	**0.91**	0.05
drinking	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05	**0.93**
Mean recognition accuracy = 93.5%												

*TFW = taking from wallet; TOW = taking out wallet; PB = packing backpacks; WB = wearing backpacks.

TABLE 4. Confusion matrix of individual HOI class over the NTU RGB+D dataset using KATH.

HOI Classes	drink water	eat meal	brush teeth	brush hair	tear up paper	put on jacket	take off jacket	put on a hat	take off a hat	phone call	play with phone	taking a selfie
drink water	**0.95**	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
eat meal	0.02	**0.96**	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
brush teeth	0.00	0.04	**0.89**	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
brush hair	0.01	0.00	0.06	**0.93**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
tear up paper	0.00	0.02	0.00	0.00	**0.97**	0.00	0.00	0.00	0.00	0.00	0.01	0.00
put on jacket	0.00	0.00	0.00	0.00	0.00	**0.94**	0.04	0.02	0.00	0.00	0.00	0.00
take off jacket	0.00	0.00	0.00	0.00	0.00	0.05	**0.95**	0.00	0.00	0.00	0.00	0.00
put on a hat	0.00	0.00	0.00	0.00	0.00	0.01	0.00	**0.95**	0.04	0.00	0.00	0.00
take off a hat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	**0.94**	0.00	0.00	0.01
phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.93**	0.05	0.02
play with phone	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03	**0.94**	0.01
taking a selfie	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.03	**0.95**
Mean recognition accuracy = 94.16%												

cricket bowling and tennis forehand. the mean specificity of this dataset is 98% which means that it can accurately reject a sample if it does not belong to a class for which it is tested for. In case of the SYSU dataset, it is observed that the mean precision, sensitivity and F1 scores are as high as 93%, 93% and 94% respectively. Less precise results are obtained with the playing phone interaction due to its 12% false positive rate. This dataset has a mean specificity rate of 99 %. From Table 7 it is inferred that the NTU RGB +D dataset has the highest precision, sensitivity, specificity and F1 scores of above 90% among all three datasets. In this dataset only the brush teeth interaction has lower than 90% sensitivity due to the lower visibility of the object and the resemblance of the action to interactions like eating meal, brush hair. Overall, it is inferred from the results of this section that the proposed methodology is an accurate HOI recognition system.

TABLE 5. Measurements of precision, sensitivity, specificity and F1 scores for the proposed method over the sports dataset.

HOI Classes	Precision	Sensitivity	Specificity	F1 score
cricket batting	0.96	0.96	0.99	0.96
cricket bowling	0.98	0.93	0.99	0.95
croquet shot	0.96	0.95	0.99	0.95
tennis forehand	0.87	0.90	0.97	0.89
tennis serve	0.97	0.92	0.99	0.94
volleyball Smash	0.84	0.91	0.96	0.88
Mean	0.93	0.92	0.98	0.92

c: COMPARISON WITH OTHER SOTA METHODS

In this section the proposed method is compared with different methodologies adopted by researchers for HOI recognition from recent years. The action recognition accuracies of each evaluated methodology are used for comparison with the proposed system. Table 8 gives the comparison of the proposed system with other SOTA systems in recent years.

TABLE 6. Measurements of precision, sensitivity, specificity and F1 score of the proposed method over the SYSU 3D HOI dataset.

Class	Precision	Sensitivity	Specificity	F1 score	Class	Precision	Sensitivity	Specificity	F1 score
sweeping	0.95	0.94	0.99	0.94	packing backpacks	0.93	0.94	0.99	0.94
mopping	0.93	0.95	0.99	0.94	wearing backpacks	0.94	0.95	0.99	0.95
Taking from wallet	0.93	0.93	0.99	0.93	Playing phone	0.88	0.89	0.98	0.89
Taking out wallet	0.98	0.91	0.99	0.94	Calling phone	0.93	0.94	0.99	0.94
Moving chair	0.95	0.98	0.99	0.97	pouring	0.90	0.91	0.99	0.91
Sitting chair	0.98	0.96	0.99	0.97	drinking	0.93	0.93	0.99	0.93
	Mean Precision = 0.93			**Mean Sensitivity = 0.93**		**Mean Specificity = 0.99**		**Mean F1 score = 0.94**	

TABLE 7. Measurements of precision, sensitivity, specificity and F1 score of the proposed method over the NTU RGB+D dataset.

Class	Precision	Sensitivity	Specificity	F1 score	Class	Precision	Sensitivity	Specificity	F1 score
drink water	0.97	0.95	0.99	0.96	take off jacket	0.96	0.95	0.99	0.95
eat meal	0.91	0.96	0.99	0.94	put on a hat	0.93	0.95	0.99	0.94
brush teeth	0.91	0.89	0.99	0.90	take off a hat	0.95	0.94	0.99	0.94
brush hair	0.92	0.93	0.99	0.93	phone call	0.96	0.93	0.99	0.94
tear up paper	0.98	0.97	0.99	0.97	play with phone	0.91	0.94	0.99	0.93
put on jacket	0.94	0.94	0.99	0.94	taking a selfie	0.96	0.95	0.99	0.95
	Mean Precision = 0.941			**Mean Sensitivity = 0.941**		**Mean Specificity = 0.99**		**Mean F1 score = 0.940**	

TABLE 8. Comparison of HOI recognition accuracy of the proposed method with other SOTA methods over the sports dataset.

Methods	Accuracy on Sports dataset (%)
Exemplar based modeling [66]	92.5
Modeling mutual context [26]	87
Weakly supervised learning HOI [67]	83
Discriminative models [68]	82.5
Proposed Method	**92.88**

TABLE 9. Comparison of hoi recognition accuracy of the proposed method with other SOTA methods over the SYSU and the NTU RGB+D dataset.

Methods	SYSU dataset (%)	NTU RGB+D dataset
Heterogeneous features +JOULE SVM [64]	84.89	-
Sparsified graph regression [69]	77.9.	87.5
Multi-modality hierarchical fusion [70]	86.89	89.70
LAFF [71]	54.2	-
Skelton-based methods [72]	-	48.9
Mobile robot platform [73]	-	75.0.
Pairwise features [74]	-	88.6
Proposed Method	**93.5**	**94.16**

In [66], a spatial and probabilistic configuration of the object is used in the form of exemplars. In [26] a mutual context of both human and object is utilized to recognize different body arts and objects. In [67] the spatial relationship between human and object is learned based on geometrical properties. A contextual relationship between postured human body parts and the object is measured in [68]. A latent structural SVM is used for learning. The recognition rate of the proposed system is 92.88% which is higher than the systems with which it was compared. Table 9 gives the comparison of the proposed system over the SYSU and the NTU RGB+D datasets. The results of the proposed system over the SYSU and NTU datasets are compared with joint heterogeneous features learning (JOULE), sparsified graph regression, multi-modality, Local Accumulative Frame Feature (LAFF), skeleton-based methods and pairwise wise features based-models. The comparison showed a higher recognition rate for the proposed system compared to the other systems.

The comparison of the recognition accuracy of the proposed method with other methods is shown in the form of a bar graph in Fig. 14.

2) PIXEL-WISE LABELING PERFORMANCE

In this experiment the performance of semantic segmentation which is implemented to label different human body parts is observed. This experiment is repeated three times

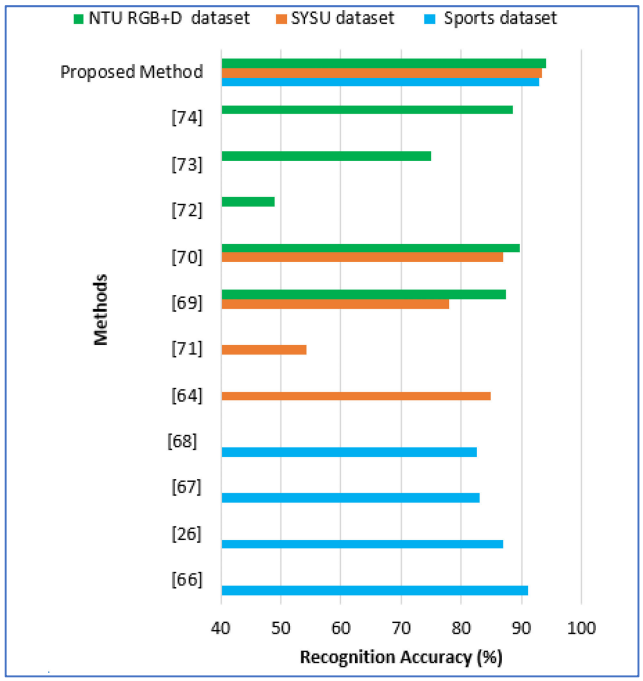


FIGURE 14. Bar graph showing comparison of the recognition accuracy of the proposed method with other methods.

to check the accuracy of pixel-level labeling in interaction classes of each dataset via CRF. For this experiment each dataset is divided into 50% for training and 50% for testing. The interaction classes for each dataset are given separately to the trained CRF model. The true positive, true negative, false positive and false negative of each labelled body part and object is evaluated individually from each class. The mean accuracy (Acc) of each labelled body part and object is calculated as;

$$Acc = \frac{(TP + TN)}{(P + N)} \tag{31}$$

The per class accuracy of each interaction class is calculated from the confusion matrix of each class individually. The accuracy measure for each body part and object of the Sports, SYSU HOI and NTU RGB+D datasets is given in Tables 10, 11 and 12 respectively. It is observed from the results of this section that the proposed technique of labeling using the elliptical model results in better labeling accuracy for each body part and object. It is also observed during experimentation that the pose of the human has effected the recognition accuracy of some body parts. For example, in the croquet interaction of the sports dataset, accuracy of neck is 79% as the person's posture is bent and the neck is not visible. Similarly, the labelling accuracy of the torso in the mopping interaction of the SYSU dataset is affected. Furthermore, the labelling accuracy of very small-sized objects such as tooth-brush and wallet is also less than for larger objects. The overall accuracy rates over three datasets are higher than 90% due to the prior phase of segmenting each body part with elliptical modeling via GMM-EM.

TABLE 10. Labelling accuracy of each body part and object over the classes of the sports dataset via CRF.

	Cbat	Cbow	Cro	TF	TS	VB	Mean per label
Head	98.2	90	88	90.9	99	92	93.02
Neck	89	92.8	79	93	85	90.2	88.17
RS	92	92	92	89	91	83	89.83
LS	90	93	94	86	93	89	90.83
RUA	94	92	91	89	87	87	90.00
LUA	92	90.4	92	91	92	90.4	91.30
Right Elbow	95	92	90	92	91	87	91.17
Left Elbow	90	92.9	89.8	94.9	89	88.7	90.88
Right Hand	93	91	92	96	89	91	92.00
Left Hand	90	89	91.6	93	87	94	90.77
Upper Torso	89.9	91	87	98	97	90.5	92.23
Lower Torso	91.2	90.9	84	99	96	93	92.35
Right Thigh	92	89	88	94	88	95	91.00
Left Thigh	90.9	87	85	86	92	98	89.82
Right Knee	92	91	96	90	93	91	92.17
Left Knee	89	93	94	94	90.9	88	91.48
Right Foot	93	92	87	91	94	92.7	91.62
Left Foot	89	94	89	95	93	95	92.50
Object	90	85	93	99	97	93	92.83
Overall label accuracy = 91.26%							

*Cbat = Cricket batting; Cbow = Cricket bowling; Cro = Croquet shot; TF = Tennis forehand; TS =Tennis serve; VB = Volleyball smash, RS=Right Shoulder, LS=Left Shoulder, RUA=Right Upper Arm, LUA= Left Upper Arm.

V. DISCUSSION

The proposed approach is tested with one outdoor RGB dataset and two indoor RGB+D datasets and showed better performance. So this system is applicable to both indoor and outdoor environments. It is a complete interaction recognition system from preprocessing of images to the recognition of each interaction class. It should be applicable to many real-world scenarios of human behavior monitoring systems, surveillance systems and smart homes, etc. However, the system also has some limitations such as the detection of smaller objects and body parts is challenging. For example, the mean labeling accuracy of the neck is 88.17%, 88.73% and 89.08% in the sports, the SYSU and the NTU RGB+D dataset respectively. This labeling accuracy is lesser than other body parts due to the occlusion of the neck by either the objects or other body parts. Similarly, in the brush teeth interaction of NTU RGB+D dataset, the labeling accuracy of the toothbrush is less than other objects due to its smaller size and occlusion by the hand of the person. However, in the proposed work we still achieved 84% labeling accuracy due to the efficient SLIC segmentation algorithm and CRF-based labeling.

TABLE 11. Labelling accuracy of each body part and object over the classes of the SYSU dataset via CRF.

	sweeping	mopping	TFW	TOW	Moving chair	Sitting chair	PB	WB	Playing phone	Calling phone	pouring	drinking	Mean per label
Head	95	92	89	92	98	97	93	92	87	89	95	87	92.17
Neck	87	89	83	90	92	91.8	92	90	81	90	95	84	88.73
RS	90	88.9	92	91	95	94	92	87	90	88	92	90	90.83
LS	86	90.5	91.9	89	97	97	90	86	91	91.5	91	92	91.08
RUA	94	93	95	89	96	97	89	92	93	86	87	88	91.58
LUA	92	93	92	91	95	94	87	91	92	91	88	91	91.42
Right Elbow	87	88	90.2	87	95	92	95	94	90.2	92	90	92	91.03
Left Elbow	89	92.5	89.9	89	94.7	93	92	96	91	94	92	94	92.26
Right Hand	90.9	92.9	94	86	92	94	90	92	87	85	81	87	89.32
Left Hand	92	94	92	89	94.9	96	92	90.3	89.4	89	83	93	91.22
Upper Torso	89	90	87	93	94	91	86	92	87	93	89	90.4	90.12
Lower Torso	86	92	90	96	92	92	84	92	85	95	84	94	90.17
Right Thigh	92	95	92	94	85	85	89	93	90	97	91.9	92	91.33
Left Thigh	94	96	90.8	97	89	87	91	90	91	95	91.6	95	92.28
Right Knee	95	91	96	92	88	91	94	91	92	93	93	96	92.67
Left Knee	95	93	94	90	90.9	93	93	89	96	92	93	93	92.66
Right Foot	96	92	90	90.3	90	96	94	92	95	90.3	92	90.9	92.38
Left Foot	94.5	94	89	91	93	95	92	90	90	92	92	92	92.04
Object	92	92	89	86	97	89	93	89	90.9	84	90	89	90.08
Overall label accuracy = 91.23%													

*RS=Right Shoulder, LS=Left Shoulder, RUA=Right Upper Arm, LUA= Left Upper Arm, TFW = taking from wallet; TOW = taking out wallet; PB = packing backpacks; WB = wearing backpacks.

TABLE 12. Labelling accuracy of each body part and object over the classes of the NTU dataset via CRF.

	drink water	eat meal	brush teeth	brush hair	tear up paper	put on jacket	take off jacket	put on hat	take off hat	phone call	play with phone	taking a selfie	Mean per label
Head	89	92	84	90	98	92	93	90	92	89	95	87	90.92
Neck	86	91	82	88	92	90	92	91	93	90	90	84	89.08
RS	92	90	89	87	95	82	81.3	83	86	88	92	90	87.94
LS	93	90.5	91	91	97	85	88	90	92	91.5	91	92	91.00
RUA	90	88	86	85	96	83	89	88	90	86	92	92	88.75
LUA	91	91	90	90	95	86	87	91	93	91	90	91	90.50
Right Elbow	92	91	87	85	92	90	95	94	90.2	92	90	92	90.85
Left Elbow	94	92.5	90.3	89	93	92	92	96	91	94	92	94	92.48
Right Hand	89	90	88	86	90	91	90	91	87	85	84	89	88.33
Left Hand	93	94	92	91	89	93	92	92	90	89	90	93	91.50
Upper Torso	90.4	92	90	93	92	94	90.5	96	92	93	89	92	91.99
Lower Torso	94	95	94	96	95	94	92	92	89	95	84	94	92.83
Right Thigh	92	93	92	94	95	95	93	93	90	97	91.9	92	93.16
Left Thigh	95	95	90.8	97	96	97	95	96	91	95	91.6	95	94.53
Right Knee	96	92	96	92	92	94	94	91	90	93	98	96	93.67
Left Knee	93	93	94	90	94	95	93	89	92	92	96	98	93.25
Right Foot	90.9	92	93	90.3	94	96	94	92	95	90.3	95	96	93.21
Left Foot	92	94	92	91	96	95	96	94	93	92	94	95	93.67
Object	90	91	84	88	95	88	89	91	90.8	86	89	91	89.40
Overall label accuracy = 91.42%													

*RS=Right Shoulder, LS=Left Shoulder, RUA=Right Upper Arm, LUA= Left Upper Arm.

VI. CONCLUSION

In this paper, we proposed a novel framework for HOI recognition. The proposed system is based on semantic human body part segmentation and feature extraction. At first an efficient silhouette segmentation of the human and object is performed via an SLIC algorithm. Then human body parts are modeled via Gaussian-based elliptical modeling and labelled at the pixel-level using CRF. Two unique semantic features, i.e., fiducial points and cloud points, are extracted. These feature descriptors are then optimized via FLDA and classified with a KATH classifier. The validity of the proposed system is proved via extensive experimentation. The experimental

section is divided into two section. At first the performance of HOI recognition is proved via accuracy, precision, sensitivity, specificity and F1 scores. The mean accuracy achieved with the Sports dataset is 92.88%, while for the SYSU dataset it is 93.5% and for the NTU RGB+D dataset it is 94.16%. Comparison with other SOTA systems showed that the proposed semantic HOI recognition system is proved to be more precise. A few instances of confusion occur between the interaction classes based on similar objects. However, the high rate of precision, sensitivity, specificity and F1 scores proved that the proposed system has high a capability of assigning accurate labels to its class and rejecting a sample if it does not belong to a specific label. In the second experiment the performance of semantic segmentation is proved via accuracy measure of human body parts and object labeling in the interaction classes of each dataset. The overall semantic segmentation accuracy of the proposed system is 91.26% with Sports dataset, 91.23% with the SYSU dataset and 91.42% with the NTU dataset. So, the proposed system is not limited to human interaction recognition, it is also applicable to other domains of computer vision such as human body parts segmentation, labelling and human pose estimation. It should be applicable to many computer-vision based applications such as healthcare monitoring, security systems and assisted living etc.

In future, we plan to investigate new features to work on multi-human and multi-object-based systems. Furthermore, we would like to work on more complex scenarios for human action recognition. We would like to increase the efficiency of labelling by applying some deep learning techniques.

REFERENCES

[1] X. Weiyao, W. Muqing, Z. Min, L. Yifeng, L. Bo, and X. Ting, "Human action recognition using multilevel depth motion maps," *IEEE Access*, vol. 7, pp. 41811–41822, 2019.

[2] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen, "Cascaded human-object interaction recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4262–4271.

[3] M. Meng, H. Drira, and J. Boonaert, "Distances evolution analysis for online and off-line human object interaction recognition," *Image Vis. Comput.*, vol. 70, pp. 32–45, Feb. 2018.

[4] G. Yu, Z. Liu, and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *Proc. ACCV*, 2014, pp. 50–65.

[5] L. Fan, W. Wang, F. Zha, and J. Yan, "Exploring new backbone and attention module for semantic segmentation in street scenes," *IEEE Access*, vol. 6, pp. 71566–71580, 2018.

[6] C. Han, Y. Duan, X. Tao, and J. Lu, "Dense convolutional networks for semantic segmentation," *IEEE Access*, vol. 7, pp. 43369–43382, 2019.

[7] 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.

[8] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3580–3589.

[9] G. Wang, P. Luo, L. Lin, and X. Wang, "Learning object interactions and descriptions for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5235–5243.

[10] H. Hwang, C. Jang, G. Park, J. Cho, and I.-J. Kim, "ElderSim: A synthetic data generation platform for human action recognition in eldercare applications," *IEEE Access*, early access, Jan. 14, 2021, doi: 10.1109/ACCESS.2021.3051842.

[11] K. H. Cheong, S. Poeschmann, J. W. Lai, J. M. Koh, U. R. Acharya, S. C. M. Yu, and K. J. W. Tang, "Practical automated video analytics for crowd monitoring and counting," *IEEE Access*, vol. 7, pp. 183252–183261, 2019.

[12] K.-P. Chou, M. Prasad, D. Wu, N. Sharma, D.-L. Li, Y.-F. Lin, M. Blumenstein, W.-C. Lin, and C.-T. Lin, "Robust feature-based automated multi-view human action recognition system," *IEEE Access*, vol. 6, pp. 15283–15296, 2018.

[13] A.-C. Popescu, I. Mocanu, and B. Cramariuc, "Fusion mechanisms for human activity recognition using automated machine learning," *IEEE Access*, vol. 8, pp. 143996–144014, 2020.

[14] Y. Fan, S. Weng, Y. Zhang, B. Shi, and Y. Zhang, "Context-aware cross-attention for skeleton-based human action recognition," *IEEE Access*, vol. 8, pp. 15280–15290, 2020.

[15] F. Rustam, A. A. Reshi, I. Ashraf, A. Mehmood, S. Ullah, D. M. Khan, and G. S. Choi, "Sensor-based human activity recognition using deep stacked multilayered perceptron model," *IEEE Access*, vol. 8, pp. 218898–218910, 2020.

[16] B. Xu, J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Interact as you intend: Intention-driven human-object interaction detection," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1423–1432, Jun. 2020.

[17] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4D human-object interactions for joint event segmentation, recognition, and object localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1165–1179, Jun. 2017.

[18] C. N. Phyo, T. T. Zin, and P. Tin, "Complex human–object interactions analyzer using a DCNN and SVM hybrid approach," *Appl. Sci.*, vol. 9, no. 9, p. 1869, May 2019.

[19] H. Liu, T.-J. Mu, and X. Huang, "Detecting human—Object interaction with multi-level pairwise feature network," *Comput. Vis. Media*, vol. 7, no. 2, pp. 229–239, Oct. 2020.

[20] 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.

[21] S. Ubalde, Z. Liu, and M. Mejail, "Detecting subtle human-object interactions using kinect," in *Proc. CIARP*, 2014, pp. 770–777.

[22] A. Nadeem, A. Jalal, and K. Kim, "Accurate physical activity recognition using multidimensional features and Markov model for smart health fitness," *Symmetry*, vol. 12, no. 11, p. 766, Oct. 2020.

[23] 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, Dec. 2019.

[24] 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.

[25] R. Al-Akam and D. Paulus, "Local feature extraction from RGB and depth videos for human action recognition," *Int. J. Mach. Learn. Comput.*, vol. 8, no. 3, pp. 274–279, Jun. 2018.

[26] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1691–1703, Sep. 2012.

[27] W. Yan, Y. Gao, and Q. Liu, "Human-object interaction recognition using multitask neural network," in *Proc. ISAS*, 2019, pp. 323–328.

[28] M. Meng, H. Drira, M. Daoudi, and J. Boonaert, "Human-object interaction recognition by learning the distances between the object and the skeleton joints," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–6.

[29] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proc. ICCV*, 2019, pp. 9468–9477.

[30] S. Qi, W. Wang, B. Jia, J. Shen, and S. C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. ECCV*, 2018, pp. 407–423.

[31] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proc. IEEE/CVF*, Jun. 2018, pp. 8359–8367.

[32] 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.

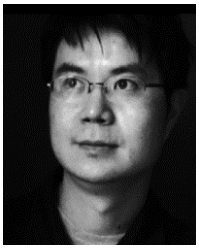
[33] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 871–885, Apr. 2019.
[34] Z. Yi, T. Chang, S. Li, R. Liu, J. Zhang, and A. Hao, "Scene-aware deep networks for semantic segmentation of images," *IEEE Access*, vol. 7, pp. 69184–69193, 2019.
[35] T. Zhou, S. Qi, W. Wang, J. Shen, and S.-C. Zhu, "Cascaded parsing of human-object interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 5, 2021, doi: 10.1109/TPAMI.2021.3049156.
[36] J. Ji, S. Buch, A. Soto, and J. C. Niebles, "End-to-end joint semantic segmentation of actors and actions in video," in *Proc. ECCV*, 2018, pp. 734–749.
[37] S. A. Khowaja and S.-L. Lee, "Semantic image networks for human action recognition," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 393–419, Oct. 2019.
[38] A. Arnab, S. Zheng, S. Jayasumana, B. Romera-Paredes, M. Larsson, and A. Kirillov, "Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 37–52, Jan. 2018.
[39] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van-Den Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 36–43.
[40] 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, pp. 271–279, Feb. 2013.
[41] K. N. Chaudhury and K. Rithwik, "Image denoising using optimally weighted bilateral filters: A sure and fast approach," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 108–112.
[42] Y. He, Y. Zheng, Y. Zhao, Y. Ren, J. Lian, and J. Gee, "Retinal image denoising via bilateral filter with a spatial kernel of optimally oriented line spread function," *Comput. Math. Methods Med.*, vol. 2017, pp. 1–13, Feb. 2017.
[43] J. Chen, Z. Li, and B. Huang, "Linear spectral clustering superpixel," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3317–3330, Jul. 2017.
[44] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[45] X. Xie, G. Xie, X. Xu, L. Cui, and J. Ren, "Automatic image segmentation with superpixels and image-level labels," *IEEE Access*, vol. 7, pp. 10999–11009, 2019.
[46] X. Xu, G. Li, G. Xie, J. Ren, and X. Xie, "Weakly supervised deep semantic segmentation using CNN and ELM with semantic candidate regions," *Complexity*, vol. 2019, Mar. 2019, Art. no. 9180391.
[47] J. Mille, A. Leborgne, and L. Tougne, "Euclidean distance-based skeletons: A few notes on average outward flux and ridgeness," *J. Math. Imag. Vis.*, vol. 61, no. 3, pp. 310–330, Jul. 2018.
[48] L. J. Latecki, Q.-N. Li, X. Bai, and W.-Y. Liu, "Skeletonization using SSM of the distance transform," in *Proc. IEEE Int. Conf. Image Process.*, Sep./Oct. 2007, pp. 349–352.
[49] T.-Q. Yan and C.-X. Zhou, "A continuous skeletonization method based on distance transform," in *Proc. ICIC*, 2012, pp. 251–258.
[50] W. Ji, X. Meng, Z. Qian, B. Xu, and D. Zhao, "Branch localization method based on the skeleton feature extraction and stereo matching for apple harvesting robot," *Int. J. Adv. Robot. Syst.*, vol. 14, no. 3, May 2017, Art. no. 172988141770527.
[51] L. Serino, C. Arcelli, and G. S. Baja, "From the zones of influence of skeleton branch points to meaningful object parts," in *Proc. DGCI*, 2013, pp. 131–142.
[52] C. Panagiotakis and A. Argyros, "Parameter-free modelling of 2D shapes with ellipses," *Pattern Recognit.*, vol. 53, pp. 259–275, May 2016.
[53] A. Arif and A. Jalal, "Automated body parts estimation and detection using salient maps and Gaussian matrix model," in *Proc. IBCAST*, 2021, pp. 667–672.
[54] T. Liu, X. Huang, and J. Ma, "Conditional random fields for image labeling," *Math. Problems Eng.*, vol. 2016, Apr. 2016, Art. no. 3846125.
[55] A. Kirillov, D. Schlesinger, W. Forkel, A. Zelenin, S. Zheng, P. H. S. Torr, and C. Rotheret, "Efficient likelihood learning of a generic CNN-CRF model for semantic segmentation," 2015, *arXiv:1511.05067v2*. [Online]. Available: https://arxiv.org/abs/1511.05067v2?source=post_page
[56] W. Zhao, Y. Fu, X. Wei, and H. Wang, "An improved image semantic segmentation method based on superpixels and conditional random fields," *Appl. Sci.*, vol. 8, no. 5, p. 837, May 2018.
[57] L. Zhang, H. Li, P. Shen, G. Zhu, J. Song, S. A. A. Shah, and M. Bennamou, "Improving semantic image segmentation with a probabilistic superpixel-based dense conditional random field," *IEEE Access*, vol. 6, pp. 15297–15310, 2018.
[58] F. Rajbdad, M. Aslam, S. Azmat, T. Ali, and S. Khattak, "Automated fiducial points detection using human body segmentation," *Arabian J. Sci. Eng.*, vol. 43, no. 2, pp. 509–524, Feb. 2018.
[59] X. Wang, H. Chen, and L. Wu, "Feature extraction of point clouds based on region clustering segmentation," *Multimedia Tools Appl.*, vol. 79, nos. 17–18, pp. 11861–11889, Jan. 2020.
[60] X.-F. Han, J. S. Jin, M.-J. Wang, W. Jiang, L. Gao, and L. Xiao, "A review of algorithms for filtering the 3D point cloud," *Signal Process., Image Commun.*, vol. 57, pp. 103–112, Sep. 2017.
[61] K. Atighehchi and R. Rolland, "Optimization of tree modes for parallel hash functions: A case study," *IEEE Trans. Comput.*, vol. 66, no. 9, pp. 1585–1598, Sep. 2017.
[62] W. Wu, B. Li, L. Chen, X. Zhu, and C. Zhang, "K-ary tree hashing for fast graph classification," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 5, pp. 936–949, May 2018.
[63] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1775–1789, Oct. 2009.
[64] J.-F. Hu, W.-S. Zheng, J.-H. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2186–2200, Nov. 2017.
[65] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.
[66] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, and T. Xiang, "Recognising human-object interaction via exemplar based modelling," in *Proc. ICCV*, 2013, pp. 3144–3151.
[67] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 601–614, Mar. 2012.
[68] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for static human-object interactions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 9–16.
[69] X. Gao, W. Hu, J. Tang, J. Liu, and Z. Guo, "Optimized skeleton-based action recognition via sparsified graph regression," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 601–610.
[70] Z. Ren, Q. Zhang, X. Gao, P. Hao, and J. Cheng, "Multi-modality learning for human action recognition," *Multimedia Tools Appl.*, vol. 80, no. 11, pp. 16185–16203, Mar. 2020.
[71] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, and J. Lai, "Real-time RGB-D activity prediction by soft regression," in *Proc. ECCV*, 2016, pp. 280–296.
[72] E. Cippitelli, E. Gambi, S. Spinsante, and F. Florez-Revuelta, "Evaluation of a skeleton-based method for human activity recognition on a large-scale RGB-D dataset," in *Proc. 2nd IET Int. Conf. Technol. Act. Assist. Living*, 2016, pp. 1–6.
[73] J. Lee and B. Ahn, "Real-time human action recognition with a low-cost RGB camera and mobile robot platform," *Sensors*, vol. 20, no. 10, p. 2886, May 2020.
[74] M. Li and H. Leung, "Multi-view depth-based pairwise feature learning for person-person interaction recognition," *Multimedia Tools Appl.*, vol. 78, no. 5, pp. 5731–5749, Mar. 2019.

NIDA KHALID received the M.S. degree in computer science from Air University, Islamabad, Pakistan. She is currently a Research Assistant with Air University. Her research interests include multimedia contents, artificial intelligence, machine learning, and computer vision.

YAZEED YASIN GHADI received the Ph.D. degree in electrical and computer engineering from Queensland University. He is currently an Assistant Professor of software engineering with Al Ain University. He was a Postdoctoral Researcher with Queensland University, before joining Al Ain University. He has published more than 25 peer-reviewed journals and conference papers and holds three pending patents. His current research interests include developing novel electro-acoustic-optic neural interfaces for large-scale high-resolution electrophysiology and distributed optogenetic stimulation. He was a recipient of several awards. His dissertation on developing novel hybrid plasmonic photonic on-chip biochemical sensors received the Sigma Xi Best Ph.D. Thesis Award.

MUNKHJARGAL GOCHOO (Member, IEEE) was born in Ulaanbaatar, Mongolia, in 1984. He received the B.S. and M.S. degrees in electronics engineering from Mongolian University of Science and Technology, in 2004 and 2005, respectively. He has completed his Ph.D. degree with the Department of Electrical Engineering, National Taipei University of Technology, Taiwan. He was with the Electronics Department, Mongolian University of Science and Technology, as a Lecturer, from 2005 to 2011. His main research interests include telecare, eldercare, the Internet of Things, machine learning, and deep learning classification algorithms.

AHMAD JALAL received the Ph.D. degree from the Department of Biomedical Engineering, Kyung Hee University, Republic of Korea. He is currently an Associate Professor with the Department of Computer Science and Engineering, Air University, Pakistan. He worked as a Postdoctoral Research Fellow at POSTECH. His research interests include multimedia contents and artificial intelligence.



KIBUM KIM (Member, IEEE) is currently an Associate Professor with the Department of Human-Computer Interaction, Hanyang University, South Korea. His research interests include the intersection of human–computer interaction, virtual reality, serious games, and artificial intelligence.