Computers, Materials & Continua
DOI: 10.32604/cmc.2022.025671
Article

Tech Science Press

An Intelligent Framework for Recognizing Social Human-Object Interactions

Mohammed Alarfaj[1], Manahil Waheed[2], Yazeed Yasin Ghadi[3], Tamara al Shloul[4], Suliman A. Alsuhibany[5], Ahmad Jalal[2] and Jeongmin Park[6,*]

[1]Department of Electrical Engineering, College of Engineering, King Faisal University, Al-Ahsa, 31982, Saudi Arabia
[2]Department of Computer Science, Air University, Islamabad, 44000, Pakistan
[3]Department of Computer Science and Software Engineering, Al Ain University, Al Ain, 15551, UAE
[4]Department of Humanities and Social Science, Al Ain University, Al Ain, 15551, UAE
[5]Department of Computer Science, College of Computer, Qassim University, Buraydah, 51452, Saudi Arabia
[6]Department of Computer Engineering, Tech University of Korea, Siheung-si, 15073, Gyeonggi-do, Korea
*Corresponding Author: Jeongmin Park. Email: jmpark@tukorea.ac.kr
Received: 01 December 2021; Accepted: 31 March 2022

Abstract: Human object interaction (HOI) recognition plays an important role in the designing of surveillance and monitoring systems for healthcare, sports, education, and public areas. It involves localizing the human and object targets and then identifying the interactions between them. However, it is a challenging task that highly depends on the extraction of robust and distinctive features from the targets and the use of fast and efficient classifiers. Hence, the proposed system offers an automated body-parts-based solution for HOI recognition. This system uses RGB (red, green, blue) images as input and segments the desired parts of the images through a segmentation technique based on the watershed algorithm. Furthermore, a convex hull-based approach for extracting key body parts has also been introduced. After identifying the key body parts, two types of features are extracted. Moreover, the entire feature vector is reduced using a dimensionality reduction technique called t-SNE (t-distributed stochastic neighbor embedding). Finally, a multinomial logistic regression classifier is utilized for identifying class labels. A large publicly available dataset, MPII (Max Planck Institute Informatics) Human Pose, has been used for system evaluation. The results prove the validity of the proposed system as it achieved 87.5% class recognition accuracy.

Keywords: Dimensionality reduction; human-object interaction; key point detection; machine learning; watershed segmentation

1 Introduction

Recognizing how humans interact with other humans and objects in their surroundings can be crucial for several applications, including crowd monitoring [1], surveillance [2], healthcare [3], sports [4], behaviour monitoring [5], smart homes [6], e-learning [7], people counting [8], human tracking [9], and event detection [10]. This makes human activity recognition (HAR) [11], human-human

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

interaction (HHI) recognition [12], and human-object interaction (HOI) recognition trending topics in computer vision and artificial intelligence (AI) fields. The scope of this research, however, is limited to HOI recognition. Although researchers have developed many well-performing HOI recognition systems in the recent past, the task remains challenging and still has room for improvement. Moreover, the use of vision [13], wearable [14], and depth sensors [15] has made it easier for modern HOI recognition systems to attain robustness and high-performance rates even with complex datasets.

This paper proposes an extensive yet reliable system for HOI recognition in social environments. The system entails that all RGB images are pre-processed using gamma correction and bilateral filtering techniques. Then, the desired silhouettes are segmented from the images. Using those silhouettes, 12 key body points are identified. This step is followed by the feature extraction phase, which involves mining two distinctive feature descriptors including ORB (oriented FAST (Features from Accelerated Segment Test) and rotated BRIEF (Binary Robust Independent Elementary Features)) and the Radon transform. Then the feature vector is reduced using the t-SNE (t-distributed stochastic neighbor embedding) dimensionality reduction technique. Finally, the interactions are classified using multinomial logistic regression. A large publicly available dataset, the MPII (Max Planck Institute Informatics) Human Pose Dataset, was used for experimentation purposes.

The main contributions of this paper include:

- Combining a commonly used super-pixel segmentation technique with a new region merging algorithm to extract accurate human silhouettes from RGB images.
- Proposing a simple yet effective way of detecting key body points on human silhouettes.
- Designing a high-performance recognition system based on the Radon transform and ORB feature descriptors.

The rest of the paper is organized as follows: Section 2 elucidates and analyses the research work relevant to the proposed system. Section 3 describes the overall methodology of the system, which also involves an extensive pre-classification process. Section 4 describes the dataset used in the proposed work and proves the robustness of the system through different experiments. Section 5 concludes the research and notes some future works.

2 Related Work

For the past few years, researchers have been working actively on HOI recognition systems [16]. Apart from RGB images, they have also investigated their systems' behaviours using depth images [17], depth videos [18], and RGB+D (red, green, blue, depth) videos [19]. In most cases, the entire images are used directly as input, and features are extracted from them. This approach is simple and especially useful when the goal is to identify multiple HOI interactions in a single image. However, another approach is to extract human and object pairs from the images and mine their features separately. Moreover, the additional steps of posture estimation [20] and human body part detection [21] for improved feature extraction [22] are also common. This approach yields more accurate results since the object and human features provide additional contextual information. Therefore, the related work can be categorized into image-based and instance-based HOI recognition systems.

2.1 Image-based HOI Recognition Systems

The image-based approach has been extensively employed in the past [23]. Jin et al. [24] have performed human-object interaction (HOI) recognition without localizing objects or identifying human poses. They have called it detection-free HOI recognition. They used a pre-trained image

encoder and LSE-Sign (log-sum-exp sign) loss function. Since they used an imbalanced dataset with multiple labels, they normalized gradients over all the available classes in a Softmax format. Moreover, Girdhar et al. [25] proposed that focusing on humans and their parts is not always useful. In some cases, using the background and context can also be helpful. Hence, they suggested the use of attention maps. Their attentional pooling module is a trainable layer that can be replaced by a pooling operation in any standard convolutional neural network (CNN).

Gkioxari et al. [26] argued that multiple cues in an image can reveal the interaction being performed. While a person's clothes and poses are important, the scene also provides an additional source of information. Hence, they exploited contextual cues to build a strong HOI recognition system. They adapted a region-based convolutional neural network (RCNN) to use more than one region for classification and called it R∗CNN. They achieved high accuracy by jointly training the action-specific models and the feature maps. Shen et al. [27] used zero-shot learning to accurately identify the relationship between a verb and an object. Their system lacked a spatial context and was tested on a simple verb-object pair. However, these methodologies were represented by complex features and the systems had very high time complexity.

2.2 Instance-based HOI Recognition Systems

Various researchers have employed the instance-based approach for identifying human-object interactions in the past few years [28]. Khalid et al. [29] incorporated semantics in scene understanding. After pre-processing the input images, the authors segmented the humans and objects involved in the given interactions. Euclidean distance transform was used to obtain the human skeleton and then human joints were extracted from the skeleton. The authors also obtained elliptical human models via GMM (Gaussian mixture model) and utilized the CRF (conditional random field) model to label the pixels of human body parts. Moreover, Yan et al. [30] proposed an HOI recognition system based on a multi-task neural network. They offered a digital glove called "WiseGlove" to detect hand motions. The system employed YOLO (you only look once) v3 to detect objects and a deep convolutional network in order to identify the interactions. For experimentation, the authors utilized both RGB and skeletal data to achieve a good recognition rate. But the dataset only had eight action classes. Moreover, their system was only able to work with a few pre-defined objects.

Wang et al. [31] proposed the use of interaction points for recognizing human-object interactions. They posed HOI as a key point detection problem. They used a fully-convolutional approach to detect the human-object interactions. The predicted interaction points were used to localize and classify the interaction. Gkioxari et al. [32] detected the human, verb, and object triplets by localizing humans through their appearance and objects through action-specific density. They used two RGB datasets to prove the validity of their system. Similarly, Li et al. [33] proposed a 3D pose estimation system and a new benchmark named "Ambiguous-HOI". They used 2D and 3D representation networks to mine features. Moreover, a cross-modal consistency task and joint learning structure were used to represent humans and objects. They performed extensive experiments on two datasets to prove the efficiency of their system. Ghadi et al. [34] used the instance-based approach to design a gait event classification system that detected events on the basis of various features including periodic and non-periodic motion, motion direction and flow, rotational angular joints, and degrees of freedom.

3 Proposed Method

This section describes the framework of the proposed HOI recognition. Fig. 1 provides an overview of the system architecture. All RGB images were pre-processed. Gamma correction was

used for image normalization and bilateral filtering for noise removal. The pre-processed images were segmented into foreground and background. The obtained foreground contains the desired human and objects. For key point detection, binary human silhouettes were employed. Then, two kinds of features were mined. From full-body silhouettes, ORB features were extracted and from key body points, Radon transforms were obtained. Both features were concatenated and further t-SNE was applied to reduce the dimensionality of the obtained feature vector. Finally, a multinomial logistic regression classifier was used for classification. The following subsections explain each stage of the framework in detail.

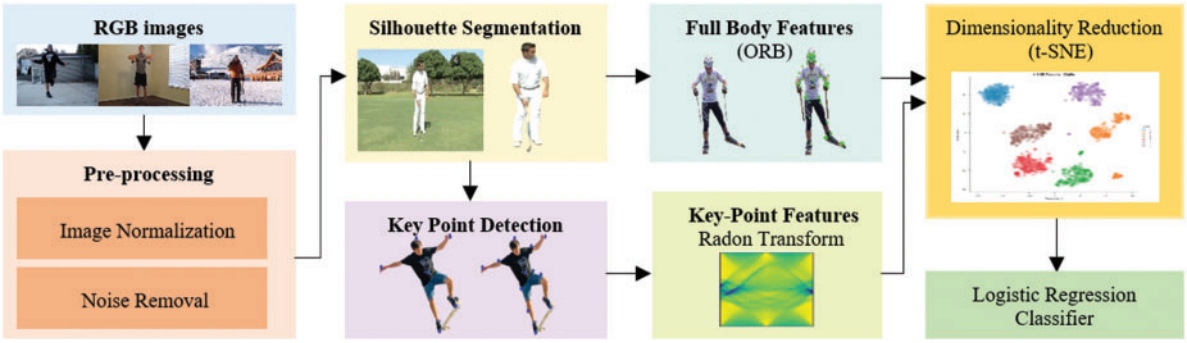
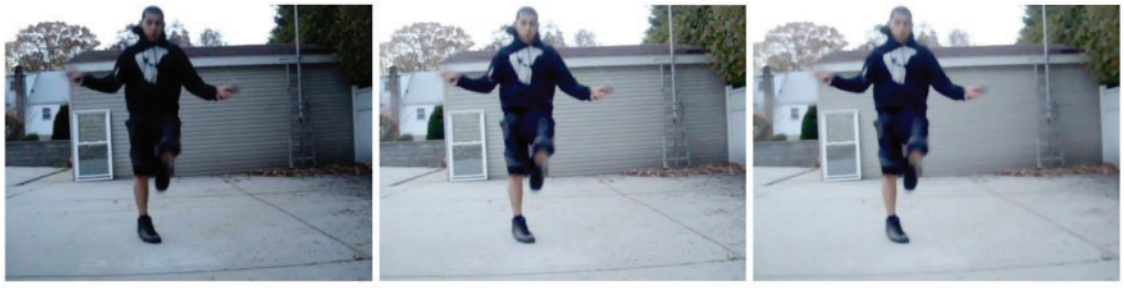


Figure 1: A general overview of the proposed HOI system

3.1 Image Pre-Processing

All RGB images in the dataset have been pre-processed using gamma correction and bilateral filtering techniques. The results of these two stages are shown in Fig. 2. Pre-processing the images at this initial stage helps in improving the overall performance of the system. Both techniques are discussed in detail in the following subsections.



(a) (b) (c)

Figure 2: Results of image pre-processing. (a) original image, (b) image after gamma correction, and (c) smooth image after applying BLF

3.1.1 Image Normalization

Image normalization is used to improve the contrast of an image by adjusting its pixel intensity values. For normalizing images in the proposed system, gamma correction has been used. This technique controls the overall brightness of an image. Several images in the used dataset look either bleached out or too dark. Gamma correction, also known as the *Power Law* transform, improves the

quality of such images. The output gamma-corrected images O have been obtained from the input images I through Eq. (1):

$$O(x) = \mathrm{I}(x_i)^{\frac{1}{G}} \tag{1}$$

where G is the gamma value that can shift the image pixels towards the higher values if set less than 1. This means that the resultant image will appear darker. However, if the gamma value is greater than 1, the image will appear lighter. A gamma value of 1 will not affect the input image.

3.1.2 Noise Removal

To improve image quality and remove noise, a bilateral filter (BLF) has been applied. A BLF smoothens the given images while preserving the edges of all the involved objects. The bilateral filter replaces the intensity value of each pixel x of the original image I with a weighted intensity value obtained from the neighboring pixels. The range kernel f_r soothes differences in the intensity values and the spatial kernel g_s flattens the differences in coordinates. The resultant smooth image I^{fil}, obtained after applying the bilateral filter, can be defined using Eq. (2).

$$I^{fil}(x) = \frac{1}{W_p} \sum_{x_i \in \Omega} I(x_i) f_r(\|I(x_i) - I(x)\|) g_s(\|x_i - x\|) \tag{2}$$

3.2 Image Segmentation

Image segmentation refers to the process of dividing an image into multiple segments, also known as super-pixels. Segmenting out various objects in an image is important for scene understanding [35]. After pre-processing the images, image segmentation has been applied and desired silhouettes have been extracted. The watershed algorithm has been used for super-pixel segmentation. It is a classic algorithm for segmentation and is especially useful when the goal is to segment touching or overlapping objects. It is a region-based method that decomposes an image completely by assigning each pixel either to a region or a watershed.

As shown in Fig. 3, the watershed algorithm divides the given image into super-pixels. To extract the desired silhouette, a super-pixel merging technique similar to the one proposed by Xu et al. [36] has been used. According to this technique, similar and adjacent super-pixels are merged to form bigger super-pixels on the basis of similarity until the desired number of super-pixels is obtained. Four types of features are extracted from each super-pixel: mean, covariance, SIFT (scale-invariant feature transform), and SURF (speeded-up robust features). The similarity of any two adjacent super-pixels is determined on the basis of the difference between the values of these features.

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

Figure 3: Results of image segmentation. (a) original image, (b) image after super-pixel segmentation, (c) merged super-pixels, and (d) extracted silhouette

3.3 Key Body Point Detection

After extracting the full-body silhouette, 12 key body points have been identified using an approach similar to the one suggested by Dargazany et al. [37] in Algorithm 1. The segmented silhouette has been converted into a binary silhouette first and then its contour has been obtained. Then a convex hull has been drawn around the contour. Points on the convex hull that were also part of the original contour have been obtained. Only five such points have been chosen. Furthermore, an additional point has been obtained by finding the centroid of the contour. Using the obtained 6 points, 6 additional key points have also been found. The method of finding these additional points is simple: the mid-point of any two key points has been found and a point on the contour lying closest to the obtained mid-point has been stored as an additional key point. Each step of the process is shown in Fig. 4.

Algorithm 1: Key Point Detection

Input: segmented silhouettes
Output: *KeyBodyPoints (kp1, kp2, kp3 . . . kp12)*
contour ← Getcontour(silhoutte)
Convex hull ← DrawConvexhull(contour)
%detecting first 5 key points%
for *point* on convexhull:
  if *point* in contour:
    KeyBodyPoints.append(point)
%detecting 6th key point%
Center ← GetContourcenter(contour)
KeyBodyPoints.append(Center)
%detecting additional 6 key points%
LE ← FindMidpoint(KeyBodyPoints [0], KeyBodyPoints [1])
lelbow ← Findclosestpointoncontour (LE)
RE ← FindMidpoint(KeyBodyPoints [2], KeyBodyPoints [1])
relbow ← Findclosestpointoncontour (RE)
LH ← FindMidpoint(KeyBodyPoints [3], KeyBodyPoints [1])
lhip ← Findclosestpointoncontour (LH)
RH ← FindMidpoint(KeyBodyPoints [4], KeyBodyPoints [1])
rhip ← Findclosestpointoncontour (RH)
LK ← FindMidpoint(KeyBodyPoints [3], KeyBodyPoints [5])
lknee ← Findclosestpointoncontour (LK)
RK ← FindMidpoint(KeyBodyPoints [4], KeyBodyPoints [5])
rknee ← Findclosestpointoncontour (RK)
KeyBodyPoints.append(lelbow, relbow, lhip, rhip, lknee, rknee)
return *KeyBodyPoints (kp1, kp2, kp3 . . . kp12)*

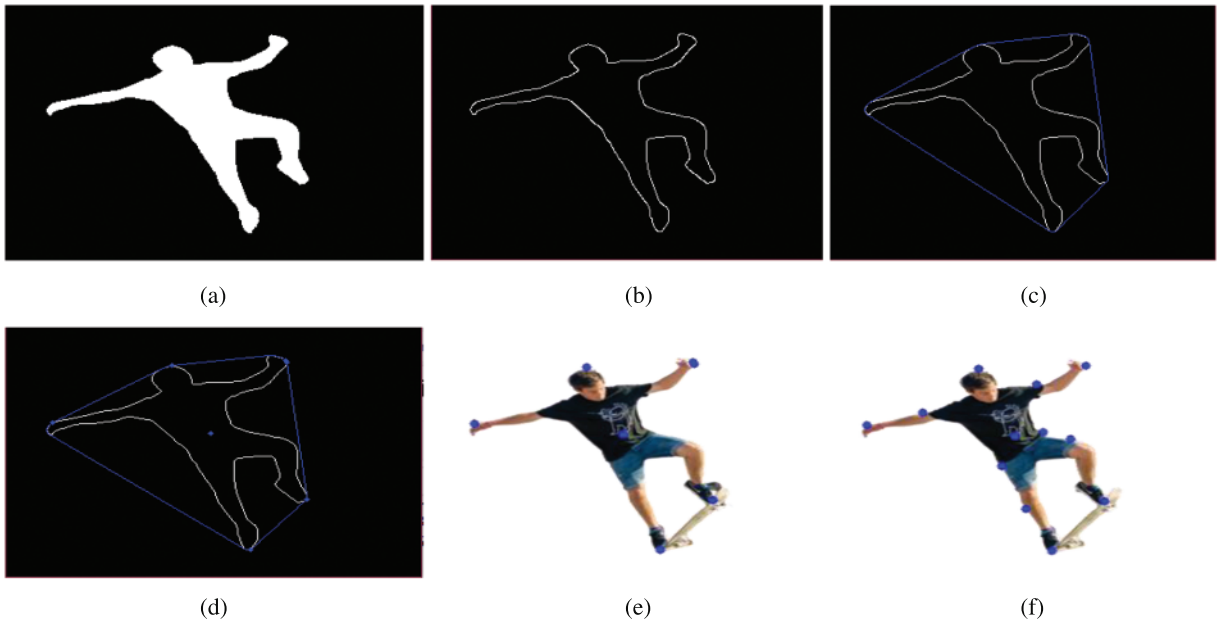


Figure 4: Results of key point detection. (a) segmented human silhouette, (b) silhouette contour, (c) convex hull around the silhouette, (d) convex hull vertices lying on the contour, (e) 6 detected key points, and (f) 12 extracted key points

3.4 Feature Extraction

The proposed system uses two types of features: ORB and Radon transform. The details and results of these features are described in the following subsections.

3.4.1 Full Body Features–ORB

Oriented FAST and rotated BRIEF (ORB) [38] is a fast and robust feature detector. It uses the FAST (Features from Accelerated Segment Test) keypoint detector for key point detection. Moreover, it is a modified version of the visual descriptor BRIEF (Binary Robust Independent Elementary Features). ORB is scale and rotation invariant. Moments of a patch can be defined using Eq. (3).

$$m_{pq} = \sum x^p y^q I\,(x, y) \tag{3}$$

where p and q are the intensity values of the image pixels at x and y locations respectively. With these moments, the centre of mass can be found using Eq. (4).

$$C = \frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \tag{4}$$

The orientation of the patch is given by Eq. (5).

$$\theta = \text{atan}\,(m_{01}, m_{10}) \tag{5}$$

Fig. 5 shows the results of applying an ORB feature descriptor to the extracted silhouettes.

(a) (b)

Figure 5: Results of applying an ORB feature descriptor. (a) skiing class–human silhouette (left) and its key feature points detected by ORB (right) and (b) golf class–human silhouette (left) and its key feature points detected by ORB (right)

3.4.2 Key-Point Features–Radon Transform

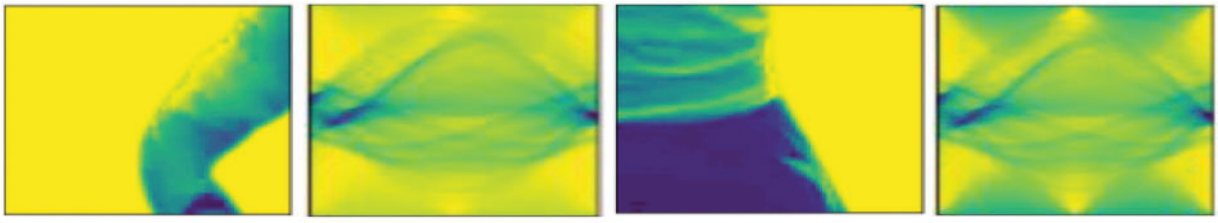
Various types of transforms have been used as feature descriptors in the past including Gabor [39] and Fourier transforms [40]. The proposed system uses the Radon transform. The Radon transform maps from the Cartesian rectangular coordinates (x, y) to polar coordinates (ρ, θ). The resulting projection is the sum of the intensities of the pixels in each direction. The Radon transform $R(\rho, \theta)$ of an image $f(x, y)$ can be obtained as follows:

$$R(\rho, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, \delta(\rho - x\cos\theta - y\sin\theta)\, \mathrm{dxdy} \tag{6}$$

where

$$\rho = x\cos\theta + y\sin\theta \tag{7}$$

Fig. 6 shows the radon transforms taken with respect to various key body points. A 30x0 window has been drawn around each key point and the radon transform of the obtained image window has been obtained.



(a) (b)

Figure 6: Radon transforms taken with respect to key points. (a) a window around the right elbow key point (left) and its radon transform (right) and (b) a window around the left hip key point (left) and its radon transform (right)

After extracting the two types of features, these are concatenated. The result is a high-dimensional feature vector. Algorithm 2 gives steps of the feature extraction and concatenation process.

Algorithm 2: Feature Extraction

Input: S: extracted silhouettes, KP: key body points
Output: combined *FeatureVector (f1, f2, f3 . . . fn)*
FeatureVector←[]
for *i* in range(len(S)):
ORB ← GetORBdescriptor(S[i])
FeatureVector.append(ORB)
  for *i* in range(len(KP)):
    Radon ← GetRadonTransform (S, i)
    FeatureVector.append(Radon)
    end
end
return FeatureVector (*f1, f2, f3 . . . fn*)

3.5 Dimensionality Reduction

After extracting the two types of features from all the images, they have been concatenated and added as descriptors of each interaction class. However, this results in a very high dimensional feature vector. Therefore, dimensionality reduction has been applied. For this purpose, the t-distributed stochastic neighbor embedding (t-SNE) technique [41] has been utilized. First, it constructs a probability distribution over pairs of high-dimensional objects. Objects similar to one another are assigned higher probabilities while lower probabilities are assigned to dissimilar objects. In this case, the density of all the points (x_j) is measured under that Gaussian distribution, which is then renormalized. This gives a set of probabilities (P_{ij}) for all the points as shown in Eq. (8).

$$p_{j|i} = \frac{exp\left(-\left|\left|x_i - x_j\right|\right|^2 / 2\sigma_i^2\right)}{\sum exp\left(-||x_i - x_k||^2 / 2\sigma_i^2\right)} \tag{8}$$

Next, it defines another probability distribution over the points in the low-dimensional map. This time, it uses a student t-distribution with one degree of freedom instead of using a Gaussian distribution. This t-distribution is also known as the Cauchy distribution, as shown in Eq. (9). In this way, the second set of probabilities (Q_{ij}) in the low dimensional space is obtained.

$$q_{ij} = \frac{\left(1 + \left|\left|y_i - y_j\right|\right|^2\right)^{-1}}{\sum \left(1 + ||y_k - y_l||^2\right)^{-1}} \tag{9}$$

Then, KL (Kullback-Liebler) divergence, given by Eq. (10), is used to measure how close or far the obtained probability distributions are. The higher the value of the KL divergence, the farther the two distributions are from one another. A KL divergence of 0 means that the two distributions in question are identical.

$$KL(P||Q) = \sum^{p_{ij}} log\frac{p_{ij}}{q_{ij}} \tag{10}$$

Finally, gradient descent is used to minimize the KL cost function. This optimization results in a map that shows the similarities between the high-dimensional inputs.

3.6 Interaction Recognition

For interaction recognition, multinomial logistic regression [42] has been employed. It is an extension of binary logistic regression that is capable of classifying more than two categories of the dependent variables. For a given set of independent variables, multinomial logistic regression predicts the probabilities of the different possible outcomes of a dependent variable that is categorically distributed. As shown in Fig. 7, it computes a value y_i for various features x_i. Then, it calculates the softmax scores of different features of a given interaction using Eq. (11).

$$S(y) = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_j}} \tag{11}$$

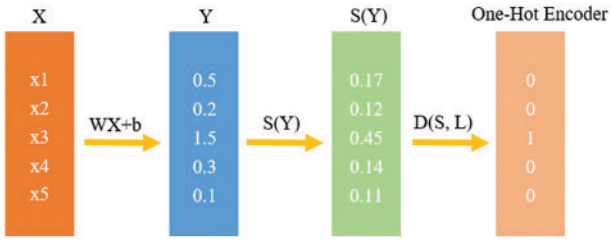


Figure 7: An overview of the layers in a multinomial logistic regression classifier

Using the softmax scores, it computes the cross-entropy loss as shown in Eq. (12).

$$D(S, L) = -\frac{1}{n}\sum_{i=1}^{n} S(y_i).log(\mathrm{S}(y_i)) \tag{12}$$

4 Performance Evaluation

This section briefly discusses the dataset that has been used for HOI recognition, the results of three different experiments that have been conducted to evaluate the proposed system, and its comparison with some recent state-of-the-art HOI recognition systems.

4.1 Dataset Description

The MPII Human Pose dataset [43] has been used to test the efficiency of the proposed system. It is a benchmark, commonly used for the evaluation of human pose estimation and is one of the most frequently used datasets for evaluating HOI recognition systems. It includes almost 25 K images containing over 40 K people performing around 410 various human activities. The activity label for each class has been provided. Images corresponding to eight of these classes are shown in Fig. 8.

Figure 8: Sample images from MPII dataset. (a) mowing lawn, (b) skiing, (c) horseback riding, (d) skateboarding, (e) bicycling, (f) golf, (g) rope skipping, and (h) ride surfboard

4.2 Experimental Settings and Results

The system has been trained and evaluated using Python (3.7) on a system having Intel Core i7 with 64-bit Windows-10. The system has a 16 GB RAM (random access memory) and 5 (GHz) CPU. To evaluate the performance of the proposed system, the LOSO (leave one subject out) cross-validation method has been employed. The performance of the proposed key-point detection technique has also been evaluated on eight randomly chosen classes of the above-mentioned dataset.

4.2.1 Experiment I: Class Recognition Accuracies

The results of classification with eight randomly chosen classes of the MPII dataset have been shown in terms of a confusion matrix in Tab. 1. It is observed during experimentation that different interactions with the same objects are often confused with each other.

Table 1: Confusion matrix showing recognition accuracies over 8 classes of MPII dataset

Classes	ML	Skiing	HR	SB	BS	Golf	RS	RD
ML	**0.88**	0.05	0	0.02	0	0	0.04	0.01
Skiing	0.04	**0.82**	0	0.05	0	0.02	0.03	0.04
HR	0	0.04	**0.89**	0.02	0.05	0	0	0
SB	0.03	0	0	**0.92**	0	0	0	0.05
BS	0	0.03	0.05	0	**0.85**	0.04	0.03	0
Golf	0.03	0.03	0	0.03	0	**0.87**	0.04	0
RS	0.04	0	0	0	0.03	0.05	**0.86**	0.02
RD	0.03	0.02	0	0.04	0	0	0	**0.91**

Mean Recognition Accuracy Rate = 87.5%

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

4.2.2 Experiment II: Precision, Recall, and F1 Scores

Tab. 2 shows the precision, recall, and F1-scores for eight classes of the MPII dataset. The results demonstrate that the proposed HOI system is able to recognize the various complex interactions with high precision. Eqs. (13)–(15) have been used to obtain the precision, sensitivity, and F1 scores of all the interaction classes for each dataset respectively.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{13}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{14}$$

$$F1\ score = \frac{2\,(Precision \times\ Recall)}{(Precision +\ Recall)} \tag{15}$$

Table 2: Precision, recall and F1 scores over 8 classes of MPII dataset

Classes	Precision	Recall	F1 score	Classes	Precision	Recall	F1 score
ML	0.89	0.88	0.88	BS	0.86	0.85	0.85
Skiing	0.83	0.82	0.82	Golf	0.85	0.87	0.86
HR	0.88	0.89	0.88	RS	0.88	0.86	0.87
SB	0.91	0.92	0.91	RD	0.92	0.91	0.91
Mean Precision= 0.877 Mean Recall= 0.875 Mean F1 Score= 0.825							

4.2.3 Experiment III: Key Point Detection Rates

This section discusses the detection rate of the key point detection algorithm proposed in the system. For this, the (probability of correct keypoint-head) PCKh@0.5 metric has been used. It is used to measure the accuracy of the localization of the body joints with the matching threshold as 50% of the head segment length. In PCKh, a joint is said to be detected correctly if the predicted joint location is within a certain threshold from the true joint location. Tab. 3 shows the PCKh scores of the 12 detected joints on 8 classes of the MPII dataset as well as the average joint detection rate.

Table 3: Key point detection rates for 8 classes of MPII dataset

Body parts	HOI Classes								
	ML	Skiing	HR	SB	BS	Golf	RS	RD	Mean
Head	95.21	92.34	89.3	92.56	98.02	96.01	93.12	91.06	93.45
Right Elbow	86.23	89.03	83.34	90.11	92.45	91.08	92.35	90.21	89.35
Left Elbow	90.29	88.09	92.12	91.34	94.62	94.12	92.2	87.02	91.23
Right Hand	86.45	90.51	91.63	88.04	96.45	97.02	90.56	86.81	90.93
Left Hand	92.34	92.16	92.21	91.57	95.23	92.75	87.14	91.66	91.88
Torso	94.32	93.12	95.06	89.86	96.23	95.12	89.03	92.32	93.13
Right Hip	87.62	88.12	90.02	87.24	94.13	92.06	95.35	93.73	91.03

(Continued)

Table 3: Continued

Body parts	HOI Classes								
	ML	Skiing	HR	SB	BS	Golf	RS	RD	Mean
Left Hip	89.32	92.15	89.09	90.13	94.27	93.03	92.42	95.72	92.02
Right Knee	90.09	92.39	94.03	86.12	92.16	94.37	90.24	92.45	91.48
Left Knee	92.43	94.23	92.09	89.25	95.09	94.45	92.76	90.03	92.54
Right Foot	89.03	90.26	87.16	93.23	94.77	90.31	86.09	91.12	90.25
Left Foot	86.26	92.15	90.9	96.46	91.03	92.04	84.03	91.32	90.52
Mean part detection rate = 91.48%									

4.2.4 Experiment IV: Comparison with Other Systems

This section shows that the proposed system performs much better than many other state-of-the-art systems. Open pose [44] is a real-time approach for 2D pose estimation of multiple people in an image. It uses part affinity fields (PAFs) to detect key human body parts. It achieves high accuracy and real-time performance regardless of the number of persons in the image. CU-Net [45] is a coupled U-Nets approach that connects semantic blocks of pairwise U-Nets so that information can flow more efficiently and feature reuse is possible across U-Net pairs. Human Pose Estimation [46] proposes a detection-followed-by-regression CNN cascade architecture that is useful for learning body part relationships and spatial context. The first part of the proposed cascade generates heat maps of part detection. The second part performs regression on these heat maps. Stacked Hourglass Networks [47] is a convolutional network architecture comprising successive pooling and up-sampling layers. It processes features across all scales and consolidates them to obtain various spatial relationships associated with the human body. Efficient FPD [48] (fast pose distillation) is a model training method for small pose CNN networks in a knowledge distillation fashion. Pose IV [49] is a single-person pose estimation technique that uses a CNN architecture based on the recently proposed Efficient Nets. It uses a multi-scale feature extractor and detection blocks based on mobile inverted bottleneck convolutions. Fig. 9 includes the pose estimation accuracies of these models on the MPII human pose dataset and proves that the proposed system outperforms them. These accuracies have been taken directly from the above-mentioned papers.

5 Discussion

The proposed system is a simple yet efficient HOI solution that can be used in real-world applications including security, surveillance, assisted living, sports, and e-learning. The system has been tested on both raw and pre-processed images and the results show that the processed images yield better results than raw images. This additional information results in accurate classification and this step also plays the most significant role in enhancing the performance of the proposed system. The evaluation section demonstrates how well the proposed system performs.

However, the proposed system has its limitations. Firstly, the system is tested only on RGB images. Evaluation of RGB+D image and video datasets can further validate the system as many researchers have achieved better results using RGB+D data. Moreover, the key point detection algorithm has its limitations, especially in images with self-occluded and deformed body parts since the proposed algorithm assumes a standard body shape while extracting the key points. Fig. 10a shows an example image in which the right foot is in front of the left thigh. Moreover, since the right foot is not in its

usual position, two key points have been detected on the left foot. This also leads to the incorrect detection of the right hip and left knee. The left hip is detected correctly but is occluded by the right foot. Fig. 10b shows an example of a human whose left hand is in front of her torso. Therefore, the left hand is not detected through the convex hull. However, the left elbow is detected instead. Since the algorithm finds the elbows by calculating the mid-points of the hands and the head, a point slightly lower than the left shoulder is detected as the left elbow in this case. Moreover, the woman's feet are hidden or missing in the original image. Therefore, the convex hull detects the two knees as the two feet. This becomes a problem when obtaining the two hip joints lying mid-way between the head and the two feet. In this case, the two hip joints are incorrectly identified on the torso and the two knee joints are identified on the hips.

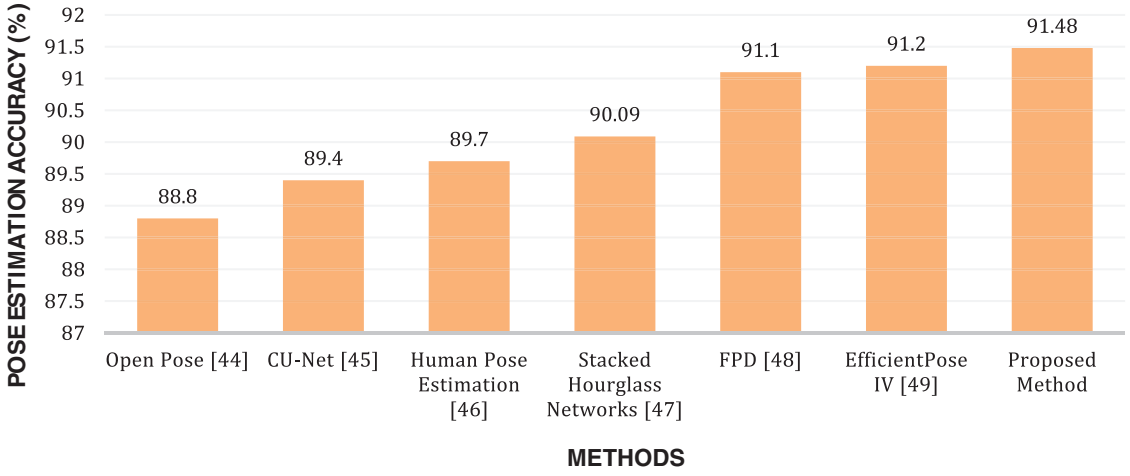
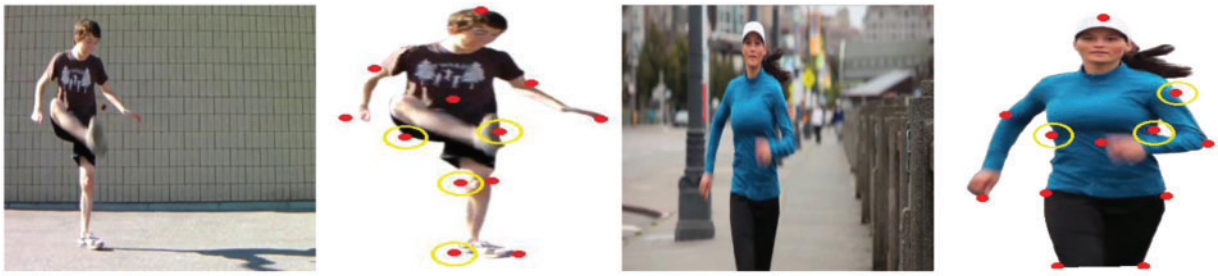


Figure 9: Comparison of mean recognition accuracy of different recent methods over HOI datasets



(a) (b)

Figure 10: Limitations of the proposed key detection algorithm. (a) Deformation and self-occlusion–original image (left) and detected key points (right) and (b) self-occlusion and hidden parts-original image (left) and detected key points (right)

6 Conclusion and Future Works

This article proposes an efficient system for the recognition of complex human-object interactions in a social environment. The RGB images are pre-processed using gamma correction and bilateral filtering techniques and then the desired human and object pairs are segmented from the images using

the watershed algorithm with an additional super-pixel merging step. Next, 12 key body points are detected on the extracted silhouette. Two kinds of features are obtained: ORB and Radon transform. A dimensionality reduction technique, t-SNE, reduces the size of the full feature vector. Finally, logistic regression classifier is used to label the interactions. The proposed system is compared with other state-of-the-art systems to demonstrate that it outperforms them. The proposed system should be applicable to many real-life applications including security systems, human-robot interactions, assisted living, and online learning.

The authors wish to work on RGBD datasets and test new and improved classifiers to achieve better performance in the future. Moreover, employing deep learning techniques is also one of our future works.

Acknowledgement: The authors acknowledge the Deanship of Scientific Research at King Faisal University for the financial support under Nasher Track (Grant No. NA000181).

Funding Statement: This research was supported by a grant (2021R1F1A1063634) of the Basic Science Research Program through the National Research Foundation (NRF) funded by the Ministry of Education, Republic of Korea.

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

References

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

[2] K. Nida, M. Gochoo, A. Jalal and K. Kim, "Modeling two-person segmentation and locomotion for stereoscopic action identification: A sustainable video surveillance system," *Sustainability*, vol. 13, no. 2, pp. 970, 2021.

[3] B. Tahir, A. Jalal and K. Kim, "IMU sensor based automatic-features descriptor for healthcare patient's daily life-log recognition," in *Proc. IBCAST*, Bhurban, Pakistan, pp. 12–16, 2021.

[4] K. Chou, M. Prasad, D. Wu, N. Sharma, D. -L. Li *et al.,* "Robust feature-based automated multi-view human action recognition system," *IEEE Access*, vol. 6, pp. 15283–15296, 2018.

[5] M. Gochoo, S. Badar, A. Jalal and K. Kim, "Monitoring real-time personal locomotion behaviors over smart indoor-outdoor environments via body-worn sensors," *IEEE Access*, vol. 9, pp. 70556–70570, 2021.

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

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

[8] M. Pervaiz, Y. Ghadi, M. Gochoo, A. Jalal, S. Kamal *et al.,* "A smart surveillance system for people counting and tracking using particle flow and modified SOM," *Sustainability*, vol. 13, no. 10, pp. 5367, 2021.

[9] M. Gochoo, S. R. Amna, G. Yazeed, A. Jalal, S. Kamal *et al.,* "A systematic deep learning based overhead tracking and counting system using RGB-D remote cameras," *Applied Sciences*, vol. 11, no. 12, pp. 5503. 2021.

[10] I. Akhter, A. Jalal and K. Kim, "Pose estimation and detection for event recognition using sense-aware features and adaboost classifier," in *Proc. IBCAST*, Bhurban, Pakistan, pp. 500–505, 2021.

[11] M. Batool, A. Jalal and K. Kim, "Sensors technologies for human activity analysis based on SVM optimized by PSO algorithm," in *Proc. ICAEM*, Taxila, Pakistan, pp. 145–150, 2019.
[12] 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 and Applications*, vol. 79, no. 11, pp. 6919–6950, 2020.
[13] H. B. Zhang, Y. X. Zhang, B. Zhong, Q. Lei, L. Yang *et al.,* "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, pp. 1–20, 2019.
[14] M. A. K. Quaid and A. Jalal, "Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm," *Multimedia Tools and Applications*, vol. 79, no. 9, pp. 6061–6083, 2019.
[15] S. Kamal and A. Jalal, "A hybrid feature extraction approach for human detection, tracking and activity recognition using depth sensors," *Arabian Journal for Science and Engineering*, vol. 41, no. 3, pp. 1043–1051, 2016.
[16] H. Liu, T. Mu and X. Huang, "Detecting human-object interaction with multi-level pairwise feature network," *Computational Visual Media*, vol. 7, no. 2, pp. 229–239, 2020.
[17] S. Kamal, A. Jalal and D. Kim, "Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM," *Journal of Electrical Engineering and Technology*, vol. 11, no. 6, pp. 1857–1862, 2016.
[18] A. Jalal, Y. -H. Kim, Y. -J. Kim, S. Kamal and D. Kim, "Robust human activity recognition from depth video using spatiotemporal multi-fused features," *Pattern Recognition*, vol. 61, pp. 295–308, 2017.
[19] A. Farooq, A. Jalal and S. Kamal, "Dense RGB-D map-based human tracking and activity recognition using skin joints features and self-organizing map," *KSII Transactions on Internet and Information Systems*, vol. 9, no. 5, pp. 1856–1869, 2015.
[20] I. Akhter, A. Jalal and K. Kim, "Adaptive pose estimation for gait event detection using context-aware model and hierarchical optimization," *Journal of Electrical Engineering and Technology*, vol. 16, no. 5, pp. 2721–2729, 2021.
[21] A. Nadeem, A. Jalal and K. Kim, "Human actions tracking and recognition based on body parts detection via artificial neural network," in *Proc. ICACS*, Lahore, Pakistan, pp. 1–6, 2020.
[22] A. Jalal and S. Kamal, "Real-time life logging via a depth silhouette-based human activity recognition system for smart home services," in *Proc. AVSS*, Seoul, South Korea, pp. 74–80, 2014.
[23] C. Phyo, T. Zin and P. Tin, "Complex human-object interactions analyzer using a DCNN and SVM hybrid approach," *Applied Sciences*, vol. 9, no. 9, pp. 1869, 2019.
[24] Y. Jin, Y. Chen, L. Wang, P. Yu, Z. Liu *et al.,* "Is object detection necessary for human-object interaction recognition?," arXiv preprint arXiv, pp. 13083, 2021.
[25] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Proc. NIPS*, Long Beach, California, USA, 2017.
[26] G. Gkioxari, R. Girshick and J. Malik, "Contextual action recognition with R∗CNN," in *Proc. ICCV*, Santiago, Chile, pp. 1080–1088, 2015.
[27] L. Shen, S. Yeung, J. Hoffman, G. Mori and L. Fei, "Scaling human-object interaction recognition through zero-shot learning," in *Proc. WACV*, Lake Tahoe, NV, USA, pp. 1568–1576, 2018.
[28] T. Zhou, W. Wang, S. Qi, H. Ling and J. Shen, "Cascaded human-object interaction recognition," in *Proc. CVPR*, virtual, pp. 4263–4272, 2020.
[29] N. Khalid, Y. Ghadi, M. Gochoo, A. Jalal and K. Kim, "Semantic recognition of human-object interactions via Gaussian-based elliptical modeling and pixel-level labeling," *IEEE Access*, vol. 9, pp. 111249–111266, 2021.
[30] W. Yan, Y. Gao and Q. Liu, "Human-object interaction recognition using multitask neural network," in *Proc. ISAS*, Albuquerque, New Mexico, pp. 323–328, 2019.
[31] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang *et al.,* "Learning human-object interaction detection using interaction points," in *Proc. CVPR*, virtual, pp. 4116–4125, 2020.

[32] G. Gkioxari, R. Girshick, P. Dollár and K. He, "Detecting and recognizing human-object interactions," in *Proc. CVPR*, Salt Lake City, Utah, pp. 8359–8367, 2018.
[33] Y. -L. Li, X. Liu, H. Lu, S. Wang, J. Liu *et al.,* "Detailed 2D-3D joint representation for human-object interaction," in *Proc. CVPR*, virtual, pp. 10163–10172, 2020.
[34] Y. Ghadi, I. Akhter, M. Alarfaj, A. Jalal and K. Kim, "Syntactic model-based human body 3D reconstruction and event classification via association based features mining and deep learning," *PeerJ Computer Science,* vol. 7, pp. 764, 2021.
[35] A. A. Rafique, A. Jalal and K. Kim, "Statistical multi-objects segmentation for indoor/outdoor scene detection and classification via depth images," in *Proc. IBCAST*, Bhurban, Pakistan, pp. 271–276, 2020.
[36] 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, no. 23, pp. 1–12, 2019.
[37] A. Dargazany and M. Nicolescu, "Human body parts tracking using torso tracking: Applications to activity recognition," in *Proc. ITNG*, Las Vegas, Nevada, pp. 646–651, 2012.
[38] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. ICCV*, Barcelona, Spain, pp. 2564–2571, 2011.
[39] M. Mahmood, A. Jalal and H. A. Evans, "Facial expression recognition in image sequences using 1D transform and gabor wavelet transform," in *Proc. ICAEM*, Taxila, Pakistan, pp. 1–6, 2018.
[40] A. Jalal, A. Ahmed, 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.
[41] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.
[42] S. P. Morgan and J. D. Teachman, "Logistic regression: Description, examples, and comparisons," *Journal of Marriage and Family*, vol. 50, no. 4, pp. 929–936, 1988.
[43] M. Andriluka, L. Pishchulin, P. Gehler and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. CVPR*, Columbus, Ohio, pp. 3686–3693, 2014.
[44] Z. Cao, G. H. Martinez, T. Simon, S. Wei and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 1, pp. 172–86, 2019.
[45] Z. Tang, X. Peng, S. Geng, Y. Zhu and D. N. Metaxas, "CU-Net: Coupled U-nets," arXiv preprint arXiv:1808.06521, 2018.
[46] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. ECCV*, Amsterdam, Netherlands, pp. 717–732, 2016.
[47] A. Newell, K. Yang and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, Amsterdam, Netherlands, pp. 483–499, 2016.
[48] F. Zhang, X. Zhu and M. Ye, "Fast human pose estimation," in *Proc. CVPR*, Long Beach, California, pp. 3517–3526, 2019.
[49] D. Groos, H. Ramampiaro and E. A. F. Ihlen, "EfficientPose: Scalable single-person pose estimation," *Applied Intelligence*, vol. 51, no. 4, pp. 2518–2533, 2020.