IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 March 2023, accepted 7 April 2023, date of publication 13 April 2023, date of current version 26 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3266774

RESEARCH ARTICLE

An Elliptical Modeling Supported System for Human Action Deep Recognition Over Aerial Surveillance

USMAN AZMAT[1], SAUD S. ALOTAIBI[2], NAIF AL MUDAWI[3], BAYAN IBRAHIMM ALABDUALLAH[4], MOHAMMED ALONAZI[5], AHMAD JALAL[1], AND JEONGMIN PARK[6]

[1]Department of Computer Science, Air University, Islamabad 44000, Pakistan
[2]Information Systems Department, Umm Al-Qura University, Makkah 24382, Saudi Arabia
[3]Department of Computer Science, College of Computer Science and Information System, Najran University, Najran 55461, Saudi Arabia
[4]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[5]Department of Information Systems, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia
[6]Department of Computer Engineering, Tech University of Korea, Siheung-si, Gyeonggi-do 15073, South Korea

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

This work was supported by the Basic Science Research Program through the National Research Foundation (NRF) (2021R1F1A1063634) funded by the Ministry of Science and Information & Communications Technology (MSIT), Republic of Korea. The authors are thankful to the Deanship of Scientific Research funded by Najran University under Research Group, Funding Program under Grant NU/RG/SERC/12/40. In addition, the authors are thankful to Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R440), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Also, the authors are thankful to Prince Satam bin Abdulaziz University by supported this study via funding from Prince Satam bin Abdulaziz University project number (PSAU/2023/R/1444).

ABSTRACT The advancement of computer vision technology has led to the development of sophisticated algorithms capable of accurately recognizing human actions from red-green-blue videos recorded by drone cameras. Hence, possessing an exceptional potential, human action recognition also faces many challenges including, tendency of humans to perform the same action in different ways, limited camera angles, and field of view. In this research article, a system has been proposed to tackle the forementioned challenges by using red-green-blue videos as input while the videos were recorded by drone cameras. First of all, the video was split into its constituent frames and then gamma correction was applied on each frame to obtain an optimized version of the image. Then the Felzenszwalb's algorithm performed the segmentation to segment out human from the input image and human silhouette was generated. Utilizing the silhouette, skeleton was extracted to spot thirteen body key points. The key points were then used to perform elliptical modeling to estimate the individual boundaries of the body parts while the elliptical modeling was governed by the Gaussian mixture model-expectation maximization algorithm. The elliptical models of the body parts were utilized to spot fiducial points that if tracked, could provide very useful information about the performed action. Some other features that were extracted for this study include, the 3d point cloud feature vector, relative distance and velocity of the key-points, and their mutual angles. The features were then forwarded for optimization under a quadratic discriminant analysis and finally, a convolutional neural network was trained to perform the action classification. Three benchmark datasets including, the Drone-Action dataset, the UAV-Human dataset, and the Okutama-Action dataset were used for a comprehensive experimentation. The system outperformed the state-of-the-art approaches by securing accuracies of 80.03%, 48.60%, and 78.01% over the Drone-Action dataset, the UAV-Human dataset, and the Okutama-Action dataset respectively.

INDEX TERMS Classification, drone, deep leaning, Felzenszwalb's segmentation, Gaussian mixture model, human action recognition.

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

I. INTRODUCTION

Artificial intelligence is playing a vital role in transforming the world into its advanced form. Numerous modern

applications pave the way to ease human life with their admirable assistance. Human Action Recognition (HAR) is one of those propitious applications that has caught and retained the attention of the researcher's community for a long time now [1], [2], [3]. There can be two major categories into which the actions of humans can be divided i.e., voluntary actions and involuntary actions. Voluntary actions may include walking, running, punching, clapping, kicking etc. while some examples of involuntary actions can be falling, limping, panicking etc. Successful recognition of human actions either voluntary or involuntary can serve in applications for elder applications [4], smart homes [5], anomaly detection [6], and life-logging [7].

Grounded or drone cameras are used to record video and then that video data is processed to recognize the action performed by the human in real-time. Working with drone cameras is more challenging as compared to grounded cameras because in the former case, the background is not static and there is more chance of noise induction into the data. Recognition of actions captured by conventional Red-Green-Blue (RGB) video cameras has been the focus of numerous previous works, e.g., [8], and [9]. Despite their usefulness, RGB data is prone to limitations such as difficulties in dealing with varying lighting conditions and cluttered backgrounds, which can pose challenges to these works. Thus, it impedes the effectiveness of the algorithm in real-world applications, such as video surveillance. Different methods for detection and representation, such as the bag of 3D points [10], skeleton joints [11], and Depth Motion Maps (DMMs) [12], [13], [14], [15], have been investigated to enhance the performance of HAR through depth images.

In their study, Shotton et al. [16] suggested an object recognition-based method that transformed pose estimation into a per-pixel classification problem, enabling the prediction of 3D positions of body joints from a single depth image. They used a large training dataset to generate 3D positions of several body joints and achieve state-of-the-art accuracy in their comparison with related work. Mathis et al. [17] proposed a method for marker less pose estimation that set its basis on deep neural networks and transfer learning. They were able to track different body parts of various species while working on a large behavior collection. They were able to score good accuracy even when only a small number of frames were labeled. In [18], a gesture-based interface for Computer-Aided Design (CAD) was introduced. The interface utilized the pose, position, velocity, and direction of fingers to perform various 3D operations, such as drawing, extruding, scaling, translating, and rotating objects. The system also featured simple binary switches that facilitated basic CAD operations while minimizing the computational cost of the system. In their work, Pareek et al. [19] employed an inertial measurement unit to extract the hand trajectory of subjects during rehabilitation tasks. The hand trajectory was then compared with the ground-truth robot trajectory to assess the real-time participation level of patients. A system introduced by Shen et al. [20] was designed to track the 3D posture of the entire arm through the use of motion and magnetic sensors on smartwatches. The system implemented a modified Hidden Markov Model (HMM) that combined data from inertial measurement unit sensors and the anatomy of arm joints to continuously estimate state variables.

The system begins by splitting the RGB video into frames then each frame is passed thorough a gamma correction phase as a preprocessing step for the image segmentation algorithm. For the action recognition, the region of interest is the one that contains the human. The Felzenszwalb's algorithm is utilized for the background rejection and extraction of the human. The Felzenszwalb's algorithm divides the whole image into small patches and then merges the correlating patches together to generate large segments that belong to a single object. Human segment, coming out as the output of the segmentation algorithm, is used to generate a black and white silhouette of the human performing the action. Afterward, a human skeleton is obtained by performing an iterative morphological erosion operation on the silhouette, which proves to be significant for extracting body key-points. A total of thirteen body key-points is extracted, consisting of the face, left upper arm, right upper arm, left lower arm, right lower arm, chest, abdomen, left thigh, right thigh, left knee, right knee, left ankle, and right ankle. An elliptical modeling based on Gaussian Mixture Model-Expectation Maximization (GMM-EM) algorithm is implemented to trace the boundaries of the body parts that are represented by the key points. The modeled ellipsoids are utilized for the extraction of the fiducial points that are tracked while the human action progresses. Moreover, the 3d point cloud feature vector, relative distance, velocity, and mutual angles of the key points are also extracted as the features of the performed action. All of the features are then optimized with the help of a Quadratic Discriminant Analysis (QDA) and sent to a Convolutional Neural Network (CNN) for classification.

The Felzenszwalb's segmentation plays a key role in this study, that's why it is essential to specifically highlight this technique. It is a highly sophisticated image segmentation algorithm [21], [22], that undertakes unsupervised approach to partition an image into regions based on similarities in color, texture, and intensity, thereby providing a highly detailed and comprehensive segmentation of the image. A significant advantage of this algorithm is its hierarchical nature, which enables it to produce a multi-scale segmentation of the image. This feature allows the algorithm to capture the finest details of the image while simultaneously identifying larger-scale structures within it. In essence, the approach represents the resulting segmentation as a tree, with each leaf node representing a single pixel, and the root node representing the entire image.

The algorithm's success lies in its innovative concept of a ''segmentation boundary''. Unlike other image segmentation algorithms that rely on clustering pixels based on their color or texture, Felzenszwalb's approach first identifies potential

boundaries between adjacent regions. It then employs a graph-based method to group these boundaries into connected components, which are subsequently interpreted as the final image segments. The robustness of the algorithm to noise and clutter in the image is another significant advantage. This is due to the fact that the algorithm does not rely on a fixed number of segments and is thus able to adapt to different levels of detail and complexity within the image. Furthermore, its hierarchical approach enables it to handle images with multiple scales of detail, making it highly suitable for tasks such as object recognition and image classification.

The major contributions of this research article are as follows:

- A novel system has been proposed that utilizes elliptical modeling based on the expectation maximization algorithm within the framework of the Gaussian mixture model.
- The implementation of state-of-the-art Felzenszwalb's algorithm, to perform efficient extraction of human subjects from their background, enabling accurate recognition of their actions.
- A 3D point cloud extraction algorithm has been presented that proved to be a very useful feature for accurate recognition of human action.
- In addition, a comprehensive comparative analysis was performed on three publicly available datasets, which feature diverse human actions. The experimental results indicate that the proposed system outperforms the state-of-the-art methods, achieving accuracies of 80.03%, 48.60%, and 78.03% on the Drone-Action dataset, the UAV-Human dataset, and the Oukutama-Action dataset, respectively.

The remainder of this paper is structured as follows. In Section II, a literature review of existing methods is provided. The proposed system is then discussed in Section III. Section IV outlines the experimental structure and presents the results obtained from the experimentation. In Section V, the performance of the proposed system over the datasets is discussed in the context of perceptions obtained from the experiments. Finally, in Section VI, conclusions are drawn, and recommendations for future research are presented.

II. RELATED WORK

Object detection [23], [24], semantic segmentation [25], [26], and CNNs [27] are key concepts that are extensively used in HAR. Object detection is a computer vision technique that involves locating and identifying objects within an image or video. This is typically done using a combination of techniques, such as feature extraction, region proposals, and classification. Object detection is widely used in applications such as autonomous driving, surveillance, and robotics. Semantic segmentation is another technique used in computer vision to segment an image into different regions based on the semantic meaning of the objects within it. This involves assigning a label to each pixel in the image, such as "sky," "tree," or "car." Semantic segmentation is commonly used in applications such as medical imaging, autonomous vehicles, and augmented reality. While CNNs are a type of deep learning algorithms that is widely used in computer vision applications. CNNs are designed to recognize patterns in images by learning from a large set of labeled data. They are highly effective at tasks such as image classification, object recognition, and face detection. Multiple methods have been proposed by different scholars for HAR utilizing a diverse range of techniques related to machine learning and deep learning. This section presents some related works on HAR systems.

A. MACHINE LEARNING-BASED HAR SYSTEMS

In this section, HAR systems that rely on machine learning approaches including supervised, unsupervised, and semi-supervised methods are presented. In [28], a unique approach focusing on Spatio-Temporal Interest Points (STIPs) was proposed for representing and recognizing human actions in video streams. For the representation of human actions, they used 2D and 3D Difference Intensity Distance Group Pattern (2D/3D-DIDGP) and employed a Support Vector Machine (SVM) for the classification. The experimentation results conclude that a 3D-DIDPG algorithm outperforms the state-of-the-art systems by a good margin but the robustness of the system could be increased by using more robust features. The proposed system adds robustness into the method with the help of features based on individual body parts. In their work, Chen et al. [29] suggested a two-level hierarchical framework for recognizing actions based on 3D skeleton data. The proposed framework aimed to overcome challenges such as high intra-class variance, movement speed variability, and high computational costs. The framework included a part-based clustering module to automatically cluster relevant joints, and a motion feature extraction and action graph module to build action graphs for recognition. The system largely depended upon the accuracy of the skeletal points. The more precise the position of the joint points, the more accurate action classification. For the precise identification of the body key points, our system uses Felzenszwalb's segmentation and extracts excellent human silhouette that results in accurate key point identification. Zhen et al. [30] implemented multiple local computation methods for HAR. The implemented methods included, STIPs, bag-of-words, sparse coding, Improved Fisher's Kernel (IFK), and Vector of Locally Aggregated Descriptors (VLAD). They also implemented naïve Bayes nearest neighbor algorithm for the classification purpose. Although this paper does not provide a complete architectural insight into the recognition of human actions, it provided a comparative analysis among various techniques, and shown that IFK technique produces the best results when working with local computation. In comparison, the proposed system provides an end-to-end architecture for the HAR and produces excellent

results. Yang et al. [31] utilized the polynormal in their work, which refers to a cluster of neighboring hypersurface normals obtained from a local spatiotemporal depth volume. Each adaptive spatiotemporal cell's low-level polynormals were aggregated by a designed scheme. The ultimate representation of depth sequences was created by concatenating the feature vectors extracted from all spatiotemporal cells. The authors implies that the features of their system could be improved by adding complementary information to them while the proposed system uses more sophisticated features that efficiently represent the actions performed by the subject. In [32], Jalal et al. extracted multi-features from the skeleton joints and used a HMM for classification purpose. They outperformed the state-of-the-art methods on the basis of their unique and compact features. Although their system performed well, but the Felzenszwalb's segmentation, that was implemented by the proposed system, generated much accurate human silhouettes, resulting in better performance as compared to them. Shahroudy et al. [33] combined RGB dense trajectories with a histogram of oriented gradients, histogram of optical flow, motion boundary histograms, and skeleton joints and used multi-class SVM for the classification. The multi-class SVM, shown comparatively lower accuracy while working with corelating actions. The proposed system handles this problem with the help of CNN-based classification. Farooq et al. [34] computed the body part of an action by using a bounding box with an optimal window size for each DMM, in order to perceive the action. The system was not able to efficiently recognize the actions in which background was coherent with the foreground. The proposed system generates excellent human silhouettes by rejecting the background using the Felzenszwalb's segmentation algorithm. A skeleton-based end-to-end model was introduced by Cui et al. [35] that enabled both person identification and action recognition. Their system relied only on skeleton-based data for both person identification and action recognition, which may limit its applicability to scenarios where only visual or other modalities of data are available. Our system begins by taking RGB video as input and performs skeleton-extraction, that enhances the applicability of the system.

B. DEEP LEARNING-BASED HAR SYSTEMS

Deep learning models are used in some HAR systems to learn features and recognize actions automatically. Zhu et al. proposed a system using co-occurrence descriptors of the skeleton-joints. The network used the trajectories of skeleton joints and a novel regularization scheme to learn co-occurrence features. A new dropout algorithm that used gates, cell, and output responses of Long-Short-Term-Memory (LSTM) neurons, was used for effective training of the network. They used three neighboring joints to represent a body part while the proposed system uses a more sophisticated approach i.e., elliptical modeling to represent the body parts [36]. Li et al. stated that long-range dynamics information is necessary and should be explicitly modeled. In response, VLAD3 was proposed as a representation that not only captured short-term dynamics using CNN but also incorporated linear dynamic systems and the VLAD descriptor to account for intermediate and long-range dynamics. Actions having short-term motion pattern, like, jumping and throwing, caused the system to be less accurate. Our system deals with this challenge by tracking the individual body parts during an action and performs better as compared to their system [37]. Shi et al. proposed to extract three stream-deep trajectory descriptors and project them on a 2D plane. They used a combination of CNN and Recurrent Neural Network (RNN) for human action recognition and classification. In their study, they do not take the complex motion of the camera in consideration that limit the performance of their framework in real-world scenarios. In the proposed system, the gamma correction of the video frames suppresses the effect of motion blur and enhances the robustness of the algorithm. [38]. Hierarchical RNN (HRNN) was used to learn temporal long-term contextual information. This method divided the human skeleton into five subparts. Each subpart was connected to one of five distinct subnetworks. HRNN was put through its paces in five different experimental settings. The system was less accurate while distinguishing between similar actions due to its high dependence on skeleton joints. Our system deals with this issue by tracking whole body parts and using 3D point cloud as an additional feature [39]. Mihanpour et al. proposed a hybrid framework consisting of CNN and a Deep Bidirectional LSTM (DB-LSTM). They used a pre-trained CNN network known as ResNet152 for the extraction of deep features from the video frames. Then these features were forwarded to a DB-LSTM for the training purpose. Their method had only been tested on a single dataset, which may limit the generalizability of the results to other datasets with different characteristics. On the contrary, we took our system through a comprehensive testing using three benchmark datasets to improve its generalizability [40]. Muhammad et al., used a combination of Bidirectional LSTM (BiLSTM) and Dilated CNN (DCNN). They selectively learned the features that had more impact on the model's accuracy with the help of DCNN. Then they forwarded those features to BiLSTM for the final classification of the human actions in the input videos. Using pre-trained weights from various AI architectures during the training stage to visually represent video frames could impact the generalizability of the results across other domains, whereas our system offers an end-to-end approach to HAR that is generalizable [41]. Kamel et al. [42] proposed an action-fusion method for HAR from depth maps and posture data using CNNs. The proposed method relied heavily on posture data descriptors to provide features for the depth maps representation. This could lead to a system failure in case of insufficient or erroneous posture data. Our proposed system, uses multiple features to tackle this issue and make more reliable classification of human actions. A deep model was proposed by Rahmani and Bennamoun [43] to effectively

model human-object interactions and intra-class variations while accounting for viewpoint changes. Model was designed to work with depth sensors, and therefore might not be applicable in scenarios where depth data is not available while our framework works on RGB video data that is common as compared to the depth data.

III. PROPOSED SYSTEM

The proposed system has been designed to analyze the input video and recognize the action performed by the human in it. First, the video is broken into frames then using the gamma correction, frames are denoised. The denoised frames are then forwarded to the segmentation block where Felzenszwalb's algorithm uses a grid graph to segment the human silhouette out of the frame. An iterative morphological erosion operation is then performed over the human silhouette to extract human skeleton out of it. Skeleton possesses the same properties as the silhouette but proves to be very useful to extract body key points. The key points are extracted from the skeleton that represent the parts of the human body and are then utilized to extract features. The features that we extracted for this research, include, the angle between the adjacent key points within a frame, the distance between the current and the previous position of the same key point while considering two consecutive frames at a time, the velocity of the key points while utilizing two consecutive frames at a time, the 3d point cloud that is extracted using the RGB image, and the fiducial points for body parts. For the fiducial points, human body parts are constructed using the elliptical modeling under the GMM-EM algorithm. All these features are then concatenated into a single data frame and labelled accordingly. After that, a QDA governed optimization process is carried out over the data. Finally, the optimized data is forwarded to train a CNN model to perform the classification. The overall architecture of the proposed system is represented by Fig. 1.

A. FRAME EXTRACTION AND GAMMA CORRECTION

A video has been provided as the input to the system but all of the analytical and computational algorithms are to be implemented over the images. That's why video has to be split into its constituent images. Gamma correction is applied on the extracted frames to denoise them [44]. It takes the image through a smoothing operation and as a result, the background of the image gets blurred and the human gets more prominent. It serves as one of the key factors on which the accuracy and efficiency of the proposed system depend. Gamma factor is used for the nonlinear relationship between the input brightness levels of an image and the perceived brightness by the human eye. It applies an inverse power-law function [45] to the pixel values of an image to make its perceived brightness linear. The power law function used for gamma correction is typically expressed as the gamma value that is a measure of the slope of the power-law curve, and the shape of curve depends upon the value of the gamma factor. When the gamma value is greater than 1, the power law curve becomes steeper, and the resulting image becomes darker and with the gamma value less than 1, the curve becomes shallower, and the resulting image becomes brighter. The governing equation for gamma correction is given below:

$$V_{OUT} = V_{IN} Gamma \tag{1}$$

where V_{OUT} is the output luminance after gamma encoding and V_{IN} is the input luminance of the pixel. Positive but less than 1 gamma values induce non-linearity in the relationship while greater than 1 value tend to make the relationship linear. In our system, we used gamma equal to 3.5. Original and the gamma corrected images are shown in Fig. 2(a) and 2(b).

B. FELZENSZWALB'S SEGMENTATION

Felzenszwalb's segmentation is a powerful image processing technique that has gained popularity in computer vision and machine learning applications. This method is widely used for image segmentation tasks, which involves partitioning an image into multiple segments or regions that share similar characteristics. The key idea behind Felzenszwalb's segmentation is to identify regions with similar colors and texture using a hierarchical grouping strategy. This results in an efficient and effective segmentation algorithm that can handle complex images with a high degree of accuracy. It uses a graph-based segmentation approach that captures visually significant regions that also have a global impact on the image. If an image is represented as a graph G with V as the set of vertices and E as the set of edges where $v_i \epsilon$ V and (v_i, v_j) ϵ E. Every edge has a weight associated with it and this weight depends upon the pixel attributes like color, motion, intensity, and location. If segmentation of the input graph G is represented by S, then S performs a division operation over G and provides G′ at the output. G′ contains distinct regions C, overall having V vertices and E′ edges while E′ $\subset$ E.

There are three important factors that are used to conclude the segmentation process.

- Intra-region difference: It is the maximum weight by which two vertices lying in the same region get connected.
- Inter-region difference: It is the minimum weight by which an edge connects two vertices lying in two different regions.
- Minimum intra-region difference between two different regions. It is given by the following equation:

$$d = \min(d_i(r_{m-1}) + \tau(r_{m-1}), d_i(r_m) + \tau(r_m)) \tag{2}$$

$$\tau(r) = k/|r| \tag{3}$$

where d is a minimum intra-region difference, d_i represents the intra-region difference, r shows the region under consideration, and k is a constant. The value of k is directly proportional to the size of the object to be segmented. Larger values of k are beneficial for the efficient segmentation of larger objects and vice versa. After the successful segmentation of the RGB image, by using basic image processing techniques, a binary human silhouette is extracted. The results for the Felzenszwalb's segmentation are represented in Fig. 3 and Algorithm 1 explains the flow of the operation.

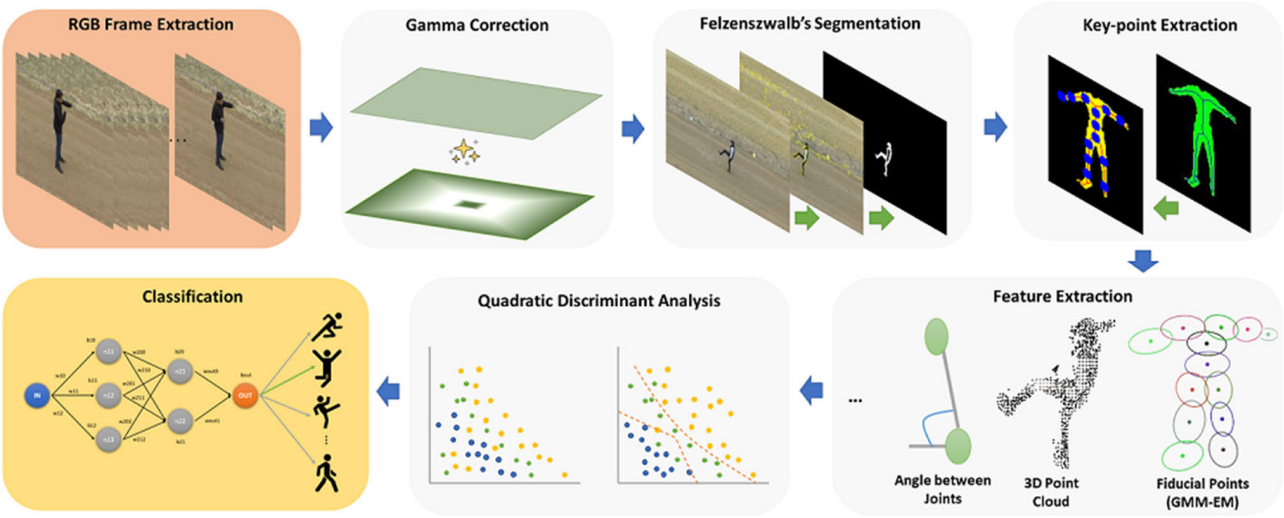

FIGURE 1. Architecture of the proposed system.

FIGURE 2. RGB Frame for (a) Hitting with stick original and gamma corrected image and (b) Waving original and gamma corrected image.

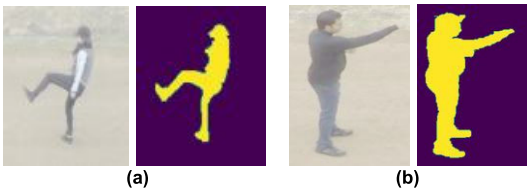

FIGURE 3. Felzenszwalb's Segmentation with (a) Kicking input and segmented image and (b) Punching input and segmented image.

Algorithm 1 Felzenszwalb's Segmentation for Silhouette Extraction

Input: RGB Video Frame
Output: Binary Human Silhouette
Method: Felzenszwalb_Segmentation (RGB image, k)
Sort (E, ascending) $= e_1, \ldots, e_m$
$R_i = v_i$
While q < m do
  If $R_i^{q-1} \neq R_j^{q-1}$ and $w\left(e_q\right) \leq d(R_i^{q-1}, R_j^{q-1})$
    $Seg^q = merge(R_i^{q-1}, R_j^{q-1})$
  Else
    $Seg^q = Seg^{q-1}$
end while
S $= Seg^m$
Sil = binarize (S, threshold)
return Sil

$^*k = constant\ ^*E = edges,\ ^*R = pixel\ regions,\ ^*Seg = segments,\ ^*d = minimum\ intra\text{-}region\ difference,\ ^*Sil = human\ silhouette$

C. SKELETONIZATION AND KEY-POINT EXTRACTION

With the help of Felzenszwalb's segmentation, the human silhouette is extracted from the RGB video frame under consideration. By performing an iterative morphological erosion operation over the binary silhouette, the skeleton is extracted which is shown in Fig. 4.

Skelton proves to be a very useful representation of the human because we can easily locate the body parts and then extract the features for the movement of each body part individually and eventually perform accurate classification of the action performed by the subject [46], [47]. The more accurate the estimated body key points, the more accurate the classification. To locate body key points, contours for the skeleton are found, and using them, a convex hull is drawn over the skeleton [48], [49]. Then extreme points of the hull

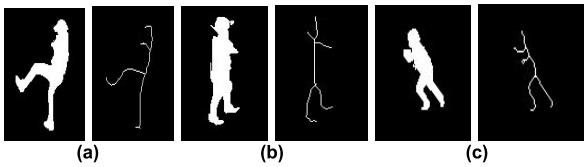

FIGURE 4. Skeletonization with (a) Kicking silhouette and skeleton (b) Hitting with bottle silhouette and skeleton and (c) jogging_side silhouette and skeleton.

are located. In this way, five key points are extracted that represent the face, the left lower arm, the right lower arm, the left ankle, and the right ankle. With the help of Eq. 4, moment M of the contours is calculated, and using Eq. 5, the x and y coordinates of the centroid of the contour are found.

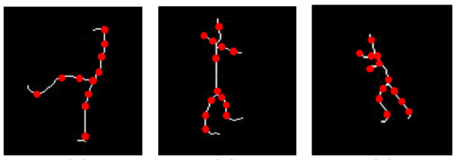

FIGURE 5. Key point extraction with (a) Kicking (b) Hitting with bottle and (c) jogging_side.

The centroid normally lies on to the lower abdomen area.

$$M_{ij} = \sum_x \sum_y x^i y^j I(x, y) \tag{4}$$

$$x, y = (M_{10}/M_{00}, M_{01}/M_{00}) \tag{5}$$

At this point, the system has computed six body key points, including, face, left lower arm, right lower arm, left ankle, right ankle, and abdomen. To locate a new key point, two most suitable neighboring points among the already calculated six points are selected and their mid-point is calculated. Afterwards, by using Euclidean distance, the nearest point of the newly found mid-point is computed that lies on the skeleton and this point is used to represent the body part. For example, for the key point of the chest, the mid-point of face and abdominal points is computed then the Euclidean distance of all the points on the skeleton with respect to the obtained mid-point is calculated. The nearest skeleton point is selected to represent the chest. Following the same procedure, points representing the left upper arm, right upper arm, left thigh, right thigh, left knee, and right knee were calculated. The formula for the computation of the mid-points is given in Eq. 6.

$$x_m, y_m = \left(\frac{x_i + x_j}{2}, \frac{y_i + y_j}{2}\right) \tag{6}$$

where x_m and y_m are the coordinates of the mid-point of two points (x_i, y_i) and (x_j, y_j). For the datasets that are used in this study, this thirteen-key-points model worked the best and produced excellent results. Visuals of key-points can be observed in Fig. 5.

D. FEATURE EXTRACTION

Features have a direct impact on the performance of a system. Good features can enhance the intelligence of the system by many folds. We extracted multiple features and by concatenating them, we generated labeled feature vectors. The feature vectors can be seen as a numerical description of the actions that are performed by the subject. Algorithm 2 describes the process of feature vector generation.

1) RELATIVE JOINT ANGLES

Relative joint angles represent the orientation of the limbs with respect to one another while performing an action. When the subject acts, the mutual angles of the limbs change with respect to one another. By keeping track of these angles, the action recognition accuracy can be increased [50]. For this purpose, the body key-points were used and a total of twelve angles were computed. The twelve specific angles for this study are shown in Fig. 6.

Algorithm 2 Feature Extraction

Input: Skeleton, body_key-points
Output: feature vector
feature_vector ← [ ]
Method: Features (Skeleton, body_key-points)
threeD_pCloud ← [ ]
fiducial_points ← [ ]
joint_angles ← [ ]
relative_distance ← [ ]
relative_velocity ← [ ]
While exit condition not true do
  threeD_pCloud ← Extract_threeD_pCloud (Skeleton, body key-points)
  fiducial_points ← Extract_ fiducial_points (Skeleton, body key-points)
  joint_angles ← Extract_ joint_angles (body_key-points)
  relative_distance ← Extract_relative_distance (body_key-points)
  relative_velocity ← Extract_ relative_velocity (body_key-points)
  feature_vector ← [threeD_pCloud, fiducial_points, joint_angles, relative_distance, relative_velocity]
end while
return feature_vector

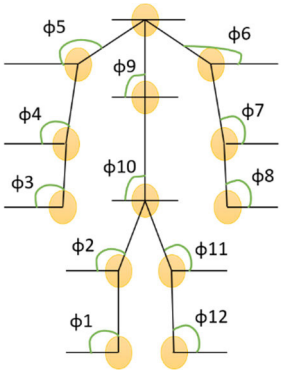

FIGURE 6. Relative joint angles for body key-points.

To calculate the angle between two points, following relation was used:

$$\varphi = tan^{-1}(y_2 - y_1/x_2 - x_1) \tag{7}$$

where (x_1, y_1) and (x_2, y_2) are the coordinates of the two points under consideration. A one-dimensional representation of the computed angles is given in Fig. 7.

2) DISTANCE AND VELOCITY

While the subject is performing an action, his relevant body parts stay in a state of continuous motion. The distance travelled by each body key point during a transition from one frame to the next and how fast the transition is performed [51], both of these factors are used as action recognition features in the proposed system. For the computation of the traveled distance, we consider two consecutive frames and find the distance between the previous and current position of each key point. By going one step ahead, we also find the rate of change of distance with respect to time to compute the velocity of every individual point. The distance and velocity

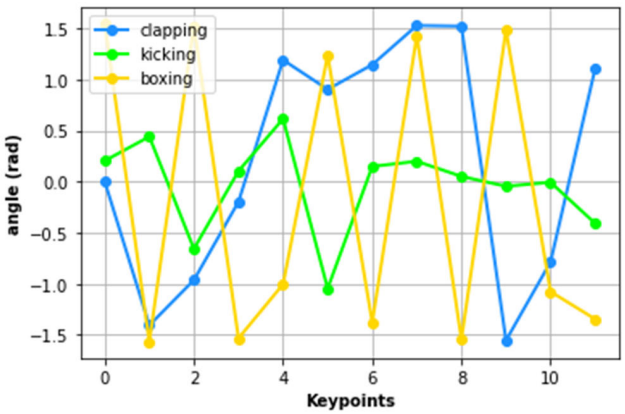

FIGURE 7. Relative joint angles for clapping, kicking, and boxing.

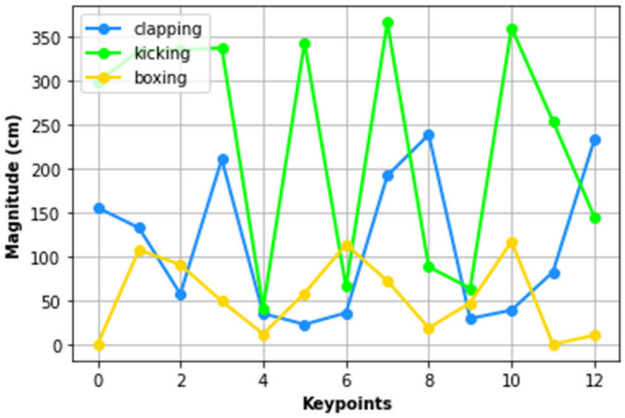

FIGURE 8. Relative distance for clapping, kicking, and boxing.

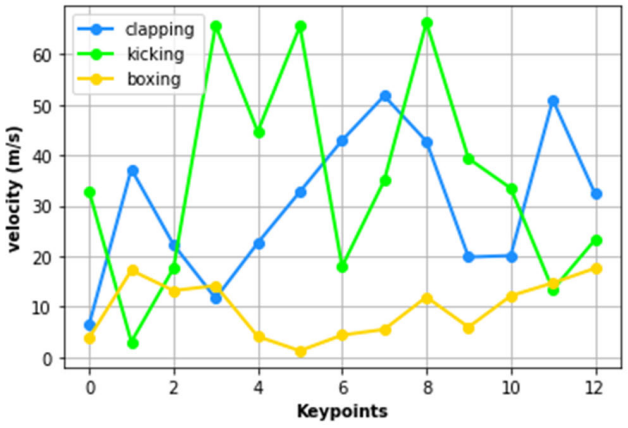

FIGURE 9. Relative velocity for clapping, kicking, and boxing.

of the key-points is represented in Fig. 8 and Fig. 9 respectively.

3) GMM-EM-BASED ELLIPTICAL MODELING

Elliptical modeling using the Expectation-Maximization (EM) algorithm is a method for fitting a Gaussian Mixture Model (GMM) to data when the covariance matrix of each component is constrained to be an elliptical shape [52], [53], [54], [55]. The EM algorithm is a two-step iterative process for finding the maximum likelihood estimates of the parameters of the GMM. The Expectation step (E-step) computes the posterior probability of each data point belonging to each component of the GMM, given the current estimates of the parameters. Mathematically, this is given by:

$$P(z_i = k|x_i, \theta) = \frac{\pi_k * N(x_i|\mu_k, \Sigma_k)}{\Sigma_j(\pi_j * N(x_i|\mu_j, \Sigma_j))} \tag{8}$$

where z_i is the latent variable indicating the elliptical component assignment for data point x_i, θ is the set of all parameters (π, μ, Σ) for the GMM, and $N(xi \mid \mu_k, \Sigma_k)$ is the probability density function of the normal distribution for k^{th} ellipse. The Maximization step (M-step) updates the estimates of the parameters by maximizing the expected complete data log-likelihood, given the posterior probabilities computed in the E-step. Mathematically, this is given by:

$$\pi_k = \frac{1}{N} * \Sigma_i(P(z_i = k|x_i, \theta)) \tag{9}$$

$$\mu_k = \frac{1}{n_k} * \Sigma_i(P\,(z_i = k \mid x_i, \theta) * x_i) \tag{10}$$

$$\Sigma_k = \frac{1}{n_k} * \Sigma_i(P\,(z_i = k \mid x_i, \theta) * (x_i - \mu_k)(x_i - \mu_k)^T) \tag{11}$$

where N is the total number of data points, n_k is the number of data points assigned to ellipse k, and the notation (x_i -μ_k)(x_i-μ_k)^T denotes the outer product of the vector difference. The E-step and M-step are repeated until convergence, which typically occurs when the change in the log-likelihood between iterations is below a certain threshold. Algorithm 3 gives complete insight into the working of the framework.

Algorithm 3 Elliptical Modeling Under GMM-EM

Input: Silhouette, Key-points
Output: elliptical model parameters (π, μ, Σ)
Method: GMM_EM(sil, kp)
$\pi \leftarrow$ mixing coefficient
$\mu \leftarrow$ means
$\Sigma \leftarrow$ covariance
While exit condition not true do
  for k in range (kp)
    $P_k \leftarrow$ Eq. 10
    $\pi_k \leftarrow$ Eq. 11
    $\mu_k \leftarrow$ Eq. 12
    $\Sigma_k \leftarrow$ Eq. 13
end while
return π, μ, Σ

For the problem under study, a binary image that consists of a human silhouette and key points are provided to the algorithm. First, circles of equal radii are drawn over the silhouette while taking the respective body key points as their centroids. The number of key points (i.e., 13) represents the number of clusters in which the whole silhouette is to be divided. After the circle assignment, the GMM-EM algorithm

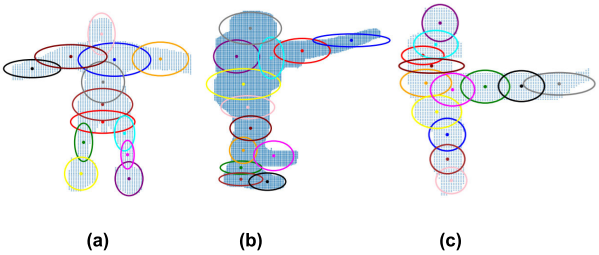

FIGURE 10. Elliptical modeling for (a) waving (b) punching and (c) Kicking.

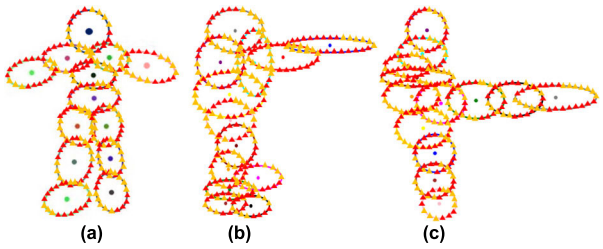

FIGURE 11. Fiducial points for (a) waving (b) punching and (c) Kicking.

starts its work and iterates over Eq. 9 through Eq. 11 and finds out the best-fit ellipses for the silhouette. In this way, we construct such a representation of the human in which each body part is represented by an independent ellipse. The elliptical modeling of the body parts proves its significance by enabling each body part to be independently trackable. Fig. 10 depicts the elliptical modeling of the actions performed.

4) FIDUCIAL POINTS

Fiducial points [56], also known as markers or landmarks, are specific points on an object or image that are used as reference points for various computer vision tasks. In HAR, fiducial points are typically used to track the movement and position of various body parts such as the face, arms, legs, knees, and ankles. By detecting these points in a sequence of images, it is possible to track the movement of a person over time, and estimate their pose, which can be used to recognize and classify different actions.

Fiducial points were spotted over the boundary of the individual body parts while utilizing the respective ellipsoids provided by the elliptical modeling phase. For this purpose, the ellipsoids were scanned in a horizontal fashion. As the internal part of the ellipsoid is dark, the transition from a high value to a low value represented a point belonging to the right boundary of the ellipsoid while a transition from a low value to high value meant the point belongs to the left boundary. Mathematically:

$$RB = rbp_1, rbp_2, \ldots, rbp_m \tag{12}$$
$$LB = lbp_1, lbp_2, \ldots, lbp_n \tag{13}$$

where RB is the right-boundary points and m represents the total number of points belonging to the right boundary. While LB and n represent the left-boundary points and the total number of points on the left-boundary respectively. After successfully separating the left and the right-boundary points, local minima and local maxima are to be located for both boundaries. Consider a point p_i (can be either from the right boundary or the left boundary), it will be a local maximum if the slope at p_i is greater than or equal to zero and the slope at p_{i+1} is less than zero. Following the same analogy, a point p_i would be a local minimum if the slope at this point is less than or equal to zero and the slope at p_{i+1} is greater than zero. Mathematically,

$$max = \{p_i | p_i' \geq 0 \; and \; p_{i+1}' < 0\} \tag{14}$$
$$min = \{p_i | p_i' \leq 0 \; and \; p_{i+1}' > 0\} \tag{15}$$

where p_i' and p_{i+1}' represent the slope of point p_i and p_{i+1} respectively. These local minima and local maxima are recorded as the fiducial points over the body part under consideration. They are stored in a vectorized form and tracked while moving from one frame to the next. We have shown the fiducial points for different actions in Fig. 11.

The red points on the boundary of ellipses represent the valley points while yellow points show the peak points. We keep track of these points over the frames and use their location information in the identification of the action.

5) 3D POINT CLOUD

A 3D point cloud is a representation of a real-world object or scene in which a set of points in 3D space are defined by their X, Y, and Z coordinates [57], [58]. These points can be used to create a detailed 3D model of an object or scene, which can be used for a variety of applications, including HAR. The proposed system generates a 3D point cloud for the human performing the action. For this purpose, we need to add an extra dimension to the pixels of the image. To add that extra dimension, we make use of the human silhouette. We first calculate the coordinates of the central pixel of the image and start iterating over the image in a horizontal fashion. We also define the focal length and scaling factor for the point cloud. By simultaneously using the RGB image and gray scale silhouette, the Z dimension is calculated by finding the intensity of the pixel that lies at the same coordinates as the RGB image. Intensity of the pixel is scaled by using the following relation:

$$Z = \frac{1}{SF} * Sil\,[u, v] \tag{16}$$

where $SF$is the scaling factor, u is the x-coordinate, and v is the y-coordinate of the pixel under consideration. The other two dimensions i.e., X and Y are computed by the following relations:

$$X = \frac{Z}{F} * (u - C_x) \tag{17}$$
$$Y = \frac{Z}{F} * (v - C_y) \tag{18}$$

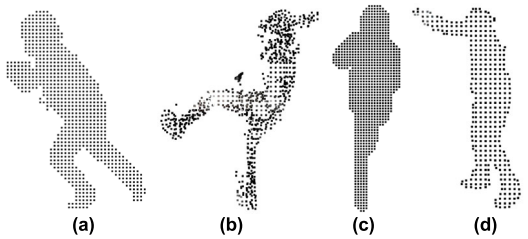

FIGURE 12. 3D point cloud for (a) jogging_side (b) kicking (c) jogging_follow (d) Stabbing.

where F is the focal length, C_x is the x-coordinate of the central pixel, and C_y is the y-coordinate of the central pixel. Using these relations, we iterate all over the pixels of the image and generate a 3D point cloud. Point clouds are very complex by nature, we can simplify them using a voxel grid filter. A voxel grid filter operates to simplify a 3D point cloud by reducing the number of points it contains. This is done by dividing the point cloud into a grid of small voxels, and then only keeping the points that fall within the voxels that have a high enough density of points. This can significantly reduce the complexity of the point cloud while still preserving its overall shape and structure. The down-sampled point cloud is then stored in a feature vector for classification. After the application of the voxel grid filter the 3D point cloud is shown in Fig. 12.

E. QUADRATIC DISCRIMINANT ANALYSIS

Quadratic discriminant analysis (QDA) is an algorithm that is used for feature optimization and classification. It utilizes a quadratic boundary to separate multiple classes. The algorithm starts by assuming that the data for each class is normally distributed with a different mean vector and covariance matrix. The QDA algorithm then calculates the likelihood of a given feature vector belonging to each class by using the normal density function. The class with the highest likelihood for the feature vector is then selected as the predicted class. To optimize the features, QDA uses a technique called regularization, which helps to prevent overfitting by adding a small positive value to the diagonal of the covariance matrix. This technique helps to make the algorithm more robust by reducing the complexity of the model [59]. In mathematical terms, we have to maximize the following ratio:

$$\frac{P(x|y=k)P(y=k)}{\Sigma P(x|y=c)P(y=c)} \tag{19}$$

where y is the class variable, k is the specific class for which the observation is being evaluated, x is the observation and c is all the possible classes. $P(x|y{=}k)$ is the class-conditional density of x given $y = k$, which is modeled as a multivariate Gaussian distribution. $P(y = k)$ is the prior probability of class k.

In our system, we have used QDA to optimize the features such that the overlap among the features reduces to lowest possible extent. A lower overlap among the features reinforces the distinctiveness of the feature vectors that represent the action performed by the subject and eventually, results in a better classification of the human actions. To witness the effect of QDA, we have displayed the original features in Fig. 13.

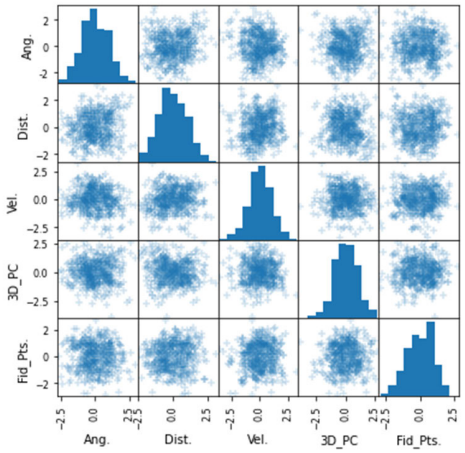

FIGURE 13. Unoptimized features extracted for HAR.

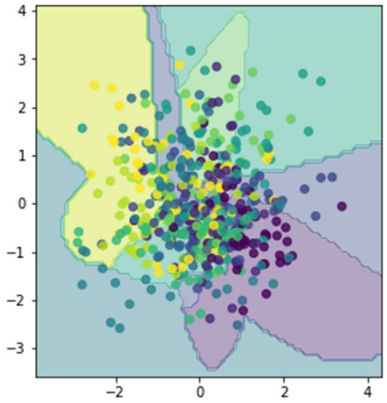

FIGURE 14. Decision boundary drawn by QDA.

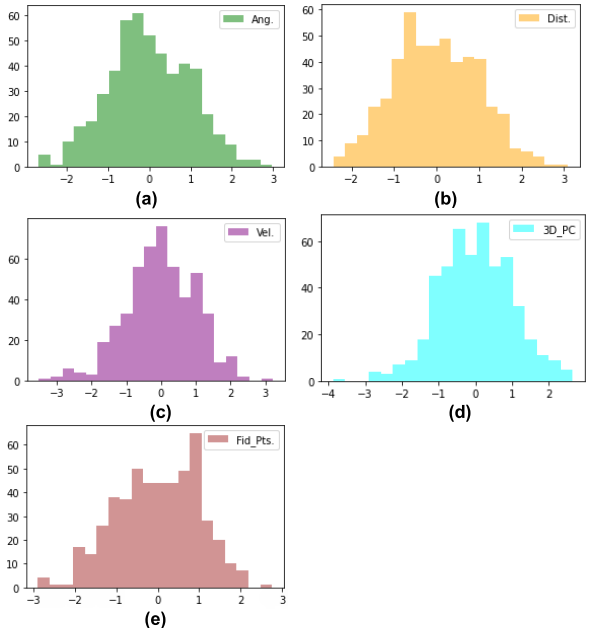

FIGURE 15. Optimized features distribution for (a) Angle (b) Distance (c) Velocity (d) 3D point cloud and (e) Fiducial Points.

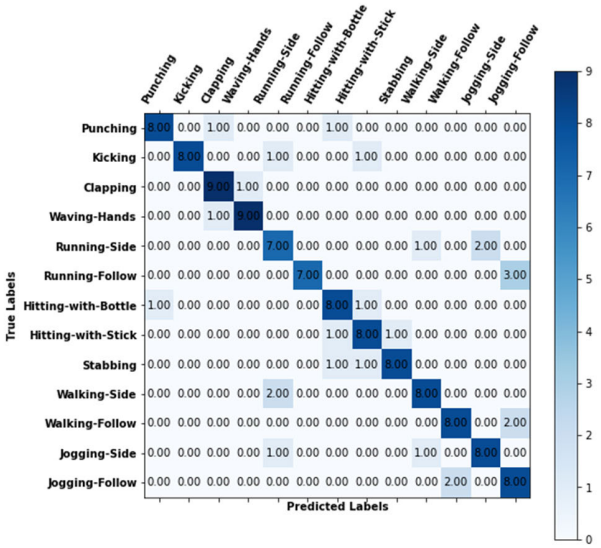

FIGURE 16. Confusion matrix for action recognition on the Drone-Action dataset using CNN.

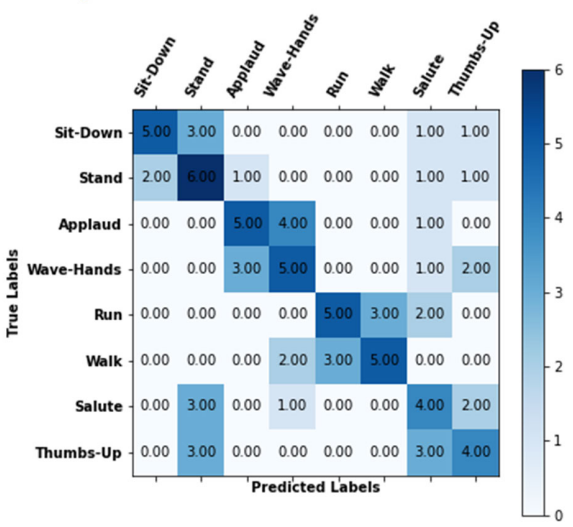

FIGURE 17. Confusion matrix for action recognition on the UAV-Human dataset using CNN.

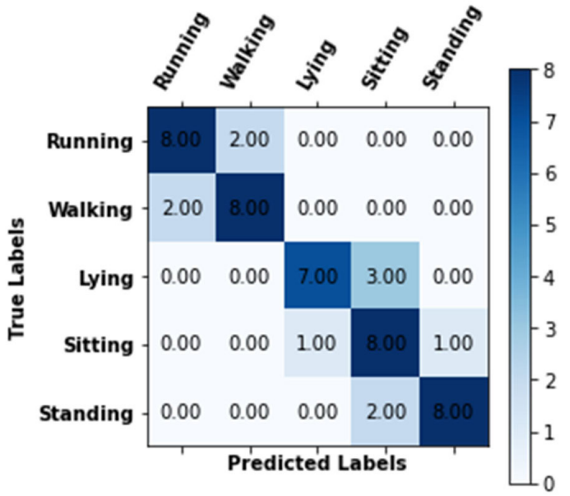

FIGURE 18. Confusion matrix for action recognition on the Okutama-Action dataset using CNN.

Fig. 14 demonstrates the decision boundaries drawn by QDA optimization algorithm and the individual distributions of the optimized features are displayed in Fig. 15.

F. HUMAN ACTION RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK

A CNN was used for the classification of the human actions [60]. Let X be the input feature map of size $H_{in}xW_{in}xC_{in}$, W be a set of learnable filters of size $KxKxC_{in}xF$, and b be a set of learnable biases of size F. Then, the output feature map Y of size $H_{out}xW_{out}xF$ is computed by the following convolution operation:

$$Y_{ijk} = \sum_{p=0}^{k-1} \sum_{q=0}^{k-1} \sum_{c=0}^{c_{in}-1} W_{pqck} X_{i+p,j+q,c} + b_k \quad (20)$$

The output of the convolution operation is then passed through a non-linear activation function like ReLU or sigmoid. Finally, a softmax layer is implemented to obtain a probability distribution over the possible classes. The equation governing softmax layer, is given below:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_i}} \quad (21)$$

where z is the previous layer's output and value of e is 2.7183. We used a blend of machine learning and deep learning by processing and optimizing the data by machine learning methods and for final classification, a CNN was utilized that resulted in excellent HAR results.

IV. EXPERIMENTAL SETTINGS AND ANALYSIS

The experimentation conducted for this research has been performed over a laptop having an Intel(R) Core (TM) i7-7500U CPU @ 2.70GHz 2.90GHz processor, 16.0 GB RAM, 64-bits Windows 10 operating system, and visual studio code as the programming tool. Furthermore, three benchmark HAR datasets namely, the Drone-Action dataset, the Okutama-Action dataset, and the UAV-Human dataset were used for this study that recorded RGB videos with the help of a drone camera from multiple angles. The 10-fold cross-validation technique was used to ensure the reliability of our research outcomes.

A. DRONE-ACTION DATASET

The Drone-Action dataset, created by Perera et al. [61] is a unique dataset that was curated specifically for HAR using footage captured by drones. It consists of 240 video clips resulting in 66919 frames. A total of 10 subjects performed 13 different actions. Some of the actions were recorded with font-view, some by side-view, and some of them were recorded while the drone followed the subject while he was performing the action. *Punching, kicking, clapping, waving hands, running-side, running-follow, hitting with a bottle,*

TABLE 1. Performance evaluation of the proposed system over drone-action dataset.

Action Classes	Precision	Recall	F1 score
Punching	0.83	0.83	0.81
Kicking	0.85	0.84	0.84
Clapping	0.90	0.88	0.88
Waving-Hands	0.87	0.85	0.84
Running-side	0.71	0.70	0.71
Running-follow	0.70	0.70	0.70
Hitting-with-Bottle	0.79	0.79	0.78
Hitting-with-Stick	0.83	0.82	0.81
Stabbing	0.80	0.79	0.78
Walking-side	0.81	0.80	0.80
Walking-follow	0.82	0.83	0.82
Jogging-side	0.81	0.80	0.79
Jogging-follow	0.80	0.82	0.80
Mean	0.81	0.80	0.80

TABLE 2. Performance evaluation of the proposed system over UAV-human dataset.

Action Classes	Precision	Recall	F1 score
Sitting	0.50	0.48	0.48
Standing	0.56	0.57	0.55
Applaud	0.53	0.52	0.50
Wave-Hands	0.52	0.53	0.52
Run	0.49	0.50	0.50
Walk	0.51	0.51	0.48
Salute	0.43	0.44	0.42
Thumbs-Up	0.40	0.40	0.37
Mean	0.49	0.49	0.48

hitting with a stick, stabbing, walking-side, walking-follow, jogging-side, and jogging-follow are the actions that were recorded by the team.

B. UAV-HUMAN DATASET

UAV-Human is a dataset [62] that covers a wide range of human actions. It provides 67,428 videos that were recorded with the contribution of 119 subjects over the span of three months. Videos were recorded in urban as well as in some rural areas also with the help of a UAV. Hence, providing a wide range of challenges in the form of different backgrounds, occlusions, weather, and camera motion. Eight human action classes from UAV-Human dataset that this study addresses to are *sit-down, stand-up, applaud, wave-hands, run, walk, thumb-up, and salute*.

TABLE 3. Performance evaluation of the proposed system over Okutama-action dataset.

Action Classes	Precision	Recall	F1 score
Running	0.82	0.82	0.82
Walking	0.80	0.79	0.77
Laying	0.69	0.71	0.70
Sitting	0.75	0.75	0.74
Standing	0.79	0.79	0.77
Mean	0.77	0.77	0.76

TABLE 4. Comparison of detection accuracy for proposed system with other state-of-the-art systems over the Drone-Action, UAV-Human, and Okutama-Action datasets.

Methods	Drone-Action (%)	UAV-Action (%)	Okutama-Action (%)
HLPF [64][61]	64.36	-	-
Pose-based CNN [65][61]	75.92	-	-
Weighted Temporal Fusion + Inception-v3+Pose-Stream [66]	78.86	-	-
3D CNN + Capsule [67]	-	-	41.87
3D CNN + BVC + Capsule [68]	-	-	47.50
Weighted Temporal Fusion + Inception-v3+Pose-Stream [66]	-	-	72.76
Baseline (SGN) [69]	-	39.99	-
MSST-RT [70]	-	41.22	-
Skel2Img [71] + DD-Net [72] + 2s-AAGCN [73] + PA-ResGCN [74] + 2s-MS-G3D [75]	-	47.44	-
Proposed system mean accuracy	80.03	48.60	78.01

C. OKUTAMA-ACTION DATASET

The Okutama-Action dataset [63] consists of 43 videos and 77365 frames. It was collected with the help of two unmanned aerial vehicles (UAVs) that were flying at a height of 10 meters to 45 meters. Some recordings were done while keeping the camera at 45 degrees and others were done at an angle of 90 degrees. For the HAR, we only considered the data that was provided for five non-interactional activities including *running, walking, lying, sitting, and standing*.

D. SYSTEM EVALUATION VIA EXPERIMENTATION

We have evaluated the proposed system for HAR on the Drone-Action dataset, the Okutama-Action dataset, and the UAV-Human dataset. To produce the most reliable results, we have repeated the experimentation three times for every dataset. Fig. 16 demonstrates the confusion matrix for the performance of the proposed system over the Drone-Action dataset. The system scores the mean accuracy of 80.03% over the mentioned dataset. While the results for the UAV-Human

dataset were produced with a mean accuracy of 48.60% and are shown in Fig. 17. The system was predicting 78.03% accurate results on average while working over the Okutama-Action dataset. The demonstration of the results can be found in Fig. 18.

Table 1 represents the proposed system's potential in the form of precision, recall, and F1-score over all classes of the Drone-Action dataset.

Table 2 displays the system's precision recall, and F1-score over all classes of the UAV-Human dataset. While the performance evaluation of the proposed system over the Okutama-Action dataset is provided in Table 3.

Finally, a performance comparison of the proposed system with state-of-the-art systems is presented in Table 4, which clearly indicates that our system achieved higher accuracy and outperformed the available state-of-the-art systems.

V. DISCUSSIONS

A HAR system was presented with gamma correction, Felzenszwalb's segmentation, GMM-EM-based elliptical modeling, multi-feature extraction, QDA-based feature optimization, and CNN-based classification as the highpoints. The system is well-suited to a wide range of real-world scenarios, including human action monitoring systems, surveillance systems, smart homes, and entertainment applications. However, it also has some limitations, such as difficulty in detecting actions in which the drone follows the subject. It can be observed from Table 1 and Table 4 that the recognition accuracy of running-side and running-follow actions was comparatively lower than the other actions. One of the reasons can be the angle by which the video is being recorded. Other reasons might include continuously varying backgrounds, occlusions, and drone movement. Due to these factors, the performance of the proposed system further drops while working on the Okutama-Action dataset, and the UAV-Human dataset. But still our system was able to beat the available state-of-the-art methods. The difference was made by the implementation of Felzenszwalb's segmentation algorithm and GMM-EM-based elliptical modeling.

VI. CONCLUSION AND FUTURE WORK

In this paper, a novel framework is proposed for human action recognition. The proposed system is based on Felzenszwalb's algorithm for segmentation and feature extraction. A unique approach that we used in our feature extraction module is to address each of the concerned body part individually with the help of GMM-EM-based elliptical modeling. The proposed system takes advantage of both machine learning and deep learning techniques to get excellent results for the algorithm. Moreover, the proposed system is not only beneficial for the recognition of human actions but can also be used for other purposes like pose estimation, and body part segmentation.

Our future plans include exploring new features for multi-human-based systems and working on more complex scenarios for human action recognition. Additionally, we aim to improve the efficiency of labelling by implementing deep learning techniques.

REFERENCES

[1] 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.

[2] 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.

[3] 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 Recognit.*, vol. 61, pp. 295–308, Jan. 2017.

[4] 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*, vol. 11, pp. 9279–9294, 2023.

[5] S. B. U. D. Tahir, A. Jalal, and M. Batool, "Wearable sensors for activity analysis using SMO-based random forest over smart home and sports datasets," in *Proc. 3rd Int. Conf. Advancements Comput. Sci. (ICACS)*, Feb. 2020, pp. 1–6.

[6] 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.

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

[8] 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.

[9] 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.

[10] R. Divya Rani and C. J. Prabhakar, "Human action recognition by concatenation of spatio-temporal 3D SIFT and CoHOG descriptors using bag of visual words," in *Proc. Int. Conf. Distrib. Comput., VLSI, Electr. Circuits Robot.*, Shivamogga, India, Oct. 2022, pp. 1–6.

[11] 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.

[12] 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.

[13] A. Jalal, J. T. Kim, and T.-S. Kim, "Human activity recognition using the labeled depth body parts information of depth silhouettes," in *Proc. 6th Int. Symp. Sustain. Healthy Buildings*, 2012, pp. 1–8.

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

[15] 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.

[16] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. CVPR*, Colorado Springs, CO, USA, Jun. 2011, pp. 1297–1304.

[17] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning," *Nature Neurosci.*, vol. 21, no. 9, pp. 1281–1289, Sep. 2018.

[18] S. Pareek, V. Sharma, and E. T. Esfahani, "Human factor study in gesture based CAD environment," in *Proc. Int. Design Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, 2015, p. 47707.

[19] S. Pareek, H. Manjunath, E. T. Esfahani, and T. Kesavadas, "MyoTrack: Realtime estimation of subject participation in robotic rehabilitation using sEMG and IMU," *IEEE Access*, vol. 7, pp. 76030–76041, 2019.

[20] S. Shen, H. Wang, and R. R. Choudhury, "I am a smartwatch and I can track my user's arm," in *Proc. 14th Annu. Int. Conf. Mobile Syst., Appl., Services*, 2016, pp. 85–96.

[21] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Proc. IEEE Int. Conf. Multimedia Expo*, Melbourne, VIC, Australia, Jul. 2012, pp. 25–30.
[22] E. C. Cahuina, J. Cousty, Y. Kenmochi, A. de Albuquerque Araújo, G. Cámara-Chávez, and S. J. F. Guimarães, "Efficient algorithms for hierarchical graph-based segmentation relying on the Felzenszwalb–Huttenlocher dissimilarity," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 33, no. 11, Oct. 2019, Art. no. 1940008.
[23] A. Ahmed, A. Jalal, and K. Kim, "Multi-objects detection and segmentation for scene understanding based on texton forest and kernel sliding perceptron," *J. Electr. Eng. Technol.*, vol. 16, no. 2, pp. 1143–1150, Mar. 2021.
[24] A. Ahmed, A. Jalal, and K. Kim, "A novel statistical method for scene classification based on multi-object categorization and logistic regression," *Sensors*, vol. 20, no. 14, p. 3871, Jul. 2020.
[25] 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.
[26] N. Khalid, 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, p. 970, Jan. 2021.
[27] S. A. Rizwan, A. Jalal, M. Gochoo, and K. Kim, "Robust active shape model via hierarchical feature extraction with SFS-optimized convolution neural network for invariant human age classification," *Electronics*, vol. 10, no. 4, p. 465, Feb. 2021.
[28] J. Arunnehru, S. Thalapathiraj, R. Dhanasekar, L. Vijayaraja, R. Kannadasan, A. A. Khan, M. A. Haq, M. Alshehri, M. I. Alwanain, and I. Keshta, "Machine vision-based human action recognition using spatio-temporal motion features (STMF) with difference intensity distance group pattern (DIDGP)," *Electronics*, vol. 11, no. 15, p. 2363, Jul. 2022.
[29] H. Chen, G. Wang, J.-H. Xue, and L. He, "A novel hierarchical framework for human action recognition," *Pattern Recognit.*, vol. 55, pp. 148–159, Jul. 2016.
[30] X. Zhen and L. Shao, "Action recognition via spatio-temporal local features: A comprehensive study," *Image Vis. Comput.*, vol. 50, pp. 1–13, Jun. 2016.
[31] X. Yang and Y. Tian, "Super normal vector for human activity recognition with depth cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 1028–1039, May 2017.
[32] A. Jalal, S. Kamal, and D. Kim, "A depth video-based human detection and activity recognition using multi-features and embedded hidden Markov models for health care monitoring systems," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 4, no. 4, p. 54, 2017.
[33] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2123–2129, Oct. 2016.
[34] A. Farooq, F. Farooq, and A. V. Le, "Human action recognition via depth maps body parts of action," *KSII Trans. Internet Inf. Syst.*, vol. 12, no. 5, pp. 2327–2347, 2018.
[35] R. Cui, G. Hua, A. Zhu, J. Wu, and H. Liu, "Hard sample mining and learning for skeleton-based human action recognition and identification," *IEEE Access*, vol. 7, pp. 8245–8257, 2019.
[36] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," 2016, *arXiv:1603.07772*.
[37] Y. Li, W. Li, V. Mahadevan, and N. Vasconcelos, "VLAD3: Encoding dynamics of deep features for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1951–1960.
[38] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1510–1520, Jul. 2017.
[39] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1110–1118.
[40] A. Mihanpour, M. J. Rashti, and S. E. Alavi, "Human action recognition in video using DB-LSTM and ResNet," in *Proc. 6th Int. Conf. Web Res. (ICWR)*, Apr. 2020, pp. 133–138.
[41] K. Muhammad, A. Ullah, A. S. Imran, M. Sajjad, M. S. Kiran, G. Sannino, and V. H. C. de Albuquerque, "Human action recognition using attention based LSTM network with dilated CNN features," *Future Gener. Comput. Syst.*, vol. 125, pp. 820–830, Dec. 2021.
[42] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 9, pp. 1806–1819, Sep. 2019.
[43] H. Rahmani and M. Bennamoun, "Learning action recognition model from depth and skeleton videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5833–5842.
[44] A. Acharya and A. V. Giri, "Contrast improvement using local gamma correction," in *Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2020, pp. 110–114.
[45] T. N. Sindhu and A. Atangana, "Reliability analysis incorporating exponentiated inverse Weibull distribution and inverse power law," *Qual. Rel. Eng. Int.*, vol. 37, no. 6, pp. 2399–2422, Oct. 2021.
[46] 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.
[47] 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.*, Mar. 2007, pp. 349–352.
[48] T.-Q. Yan and C.-X. Zhou, "A continuous skeletonization method based on distance transform," in *Proc. ICIC*, 2012, pp. 251–258.
[49] 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.
[50] M. Pervaiz, I. Akhter, and S. A. Chelloug, "An optimized system for human behaviour analysis in E-learning," in *Proc. Int. Conf. Electr. Eng. Sustain. Technol. (ICEEST)*, Dec. 2022, pp. 1–5.
[51] H. Xing and D. Burschka, "Skeletal human action recognition using hybrid attention based graph convolutional network," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 3333–3340.
[52] 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.
[53] C. Panagiotakis and A. Argyros, "Parameter-free modelling of 2D shapes with ellipses," *Pattern Recognit.*, vol. 53, pp. 259–275, May 2016.
[54] A. Arif and A. Jalal, "Automated body parts estimation and detection using salient maps and Gaussian matrix model," in *Proc. Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2021, pp. 667–672.
[55] N. Khalid, Y. 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.
[56] 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.
[57] 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.
[58] 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.
[59] A. S. Manikandan and M. Saravanan, "Application of multi-domain feature for automated seizure detection from EEG signal," in *Proc. 3rd Int. Conf. Smart Electron. Commun. (ICOSEC)*, Trichy, India, Oct. 2022, pp. 280–285.
[60] M. Kanthi, T. H. Sarma, and C. S. Bindu, "A 3D-deep CNN based feature extraction and hyperspectral image classification," in *Proc. IEEE India Geosci. Remote Sens. Symp. (InGARSS)*, Ahmedabad, India, Dec. 2020, pp. 229–232.
[61] A. G. Perera, Y. W. Law, and J. Chahl, "Drone-action: An outdoor recorded drone video dataset for action recognition," *Drones*, vol. 3, no. 4, p. 82, Nov. 2019.
[62] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16261–16270.
[63] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2153–2160.
[64] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.

[65] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3218–3226.
[66] S. K. Yadav, E. Pahwa, A. Luthra, K. Tiwari, H. M. Pandey, and P. Corcoran, "SWTF: Sparse weighted temporal fusion for drone-based activity recognition," 2022, *arXiv:2211.05531*.
[67] P. Z. Hang, P. Wei, and S. Han, "CapsNets algorithm," *J. Phys., Conf. Ser.*, vol. 1544, May 2020, Art. no. 012030.
[68] A. M. Algamdi, V. Sanchez, and C.-T. Li, "DroneCaps: Recognition of human actions in drone videos using capsule networks with binary volume comparisons," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3174–3178.
[69] L. Xu, C. Lan, W. Zeng, and C. Lu, "Skeleton-based mutually assisted interacted object localization and human action recognition," *IEEE Trans. Multimedia*, early access, Mar. 16, 2022, doi: 10.1109/TMM.2022.3175374.
[70] Y. Sun, Y. Shen, and L. Ma, "MSST-RT: Multi-stream spatial–temporal relative transformer for skeleton-based action recognition," *Sensors*, vol. 21, no. 16, p. 5339, Aug. 2021.
[71] V.-N. Hoang, T.-L. Le, T.-H. Tran, and V.-T. Nguyen, "3D skeleton-based action recognition with convolutional neural networks," in *Proc. Int. Conf. Multimedia Anal. Pattern Recognit. (MAPR)*, May 2019, pp. 1–6.
[72] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proc. ACM multimedia Asia*, 2019, pp. 1–6.
[73] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 9532–9545, 2020.
[74] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1625–1633.
[75] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 140–149.

USMAN AZMAT received the B.E. degree in mechatronics and control engineering from the University of Engineering and Technology Lahore, Lahore, Pakistan. He is currently pursuing the M.S. degree in artificial intelligence with Air University, Islamabad, Pakistan. He was a microcontroller programmer, from 2015 to 2020, and he has been a Research Associate with Air University, since 2020. His research interests include artificial intelligence, machine learning algorithms, deep learning classification, human locomotion analysis, inertial signals filtration, image and video processing, human action, and interaction recognition.

SAUD S. ALOTAIBI received the bachelor's degree in computer science from King Abdul Aziz University, in 2000, the master's degree in computer science from King Fahd University, Dhahran, in May 2008, and the Ph.D. degree in computer science from Colorado State University, Fort Collins, USA, in August 2015, under the supervision of Dr. Charles Anderson. He started his career as an Assistant Lecturer with Umm Al-Qura University, Makkah, Saudi Arabia, in July 2001. From 2015 to 2018, he was with the Deanship of Information Technology to improve the IT services that are provided to Umm Al-Qura University. After that, he was a Deputy of the IT-Center for E-Government and Application Services, Umm Al-Qura University, in January 2009, where he is currently an Assistant Professor in computer science. He is also the Vice Dean for Academic Affairs with the Computer and Information College. His current research interests include AI, machine learning, natural language processing, the neural computing IoT, knowledge representation, smart cities, wireless, and sensors.

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

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

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

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

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